SWEEP: ACHIEVING GLOBAL CONVERGENCE AND ENABLING PARETO FRONT EXPLORATION FOR DEEP NEURAL MORL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective reinforcement learning (MORL) has gained significant attention in recent years due to its wide range of applications. However, theoretical understanding on MORL algorithmic design remains in its infancy thus far, especially for the widely used deep neural network (DNN)-based actor–critic methods. This motivates us to study MORL from a theoretical perspective and develop a DNN-based actor–critic approach that (i) offers global convergence guarantees to weak Pareto-optimal policies and (ii) enables systematic exploration of the entire Pareto front (PF). Inspired by recent advances in multi-objective optimization (MOO) theory, our basic idea to enable systematic PF exploration is to scalarize the original vector-valued MORL problem with the weighted-Chebyshev (WC) technique and leverage the one-to-one correspondence between the PF and the WC-scalarization. However, the non-smoothness incurred by the WC-scalarization introduces difficulty in multi-objective policy gradient evaluation in MORL. Toward this end, we propose a new parameterized smoothed-WC actor-critic technique to overcome this aforementioned challenge to solve the deep-neural MORL problems with a global convergence rate of $\mathcal{O}(1/T)$, where T denotes the total number of iterations. Notably, to establish the global convergence rate guarantee for DNN-based MORL, we propose a new set of theoretical assumptions and analysis. To our knowledge, these assumptions and analysis are new in the literature and could be of independent interest to the field of actor-critic MORL with nonlinear approximation. Finally, extensive numerical experiments further validate the effectiveness of our method.

1 INTRODUCTION

1) Background and Motivations: Although recent years have witnessed remarkable advances in reinforcement learning (RL) (Kaelbling et al., 1996; Sutton et al., 1998; Arulkumaran et al., 2017), the traditional RL framework fails to model many complex real-world learning scenarios where multiple conflicting objectives are involved. Take reinforcement learning with human feedback (RLHF) as an example: multiple human-aligned metrics, such as *helpfulness*, *verbosity*, and *toxicity*, may conflict with each other (Ouyang et al., 2022; Wang et al., 2023; Chakraborty et al., 2024), making it insufficient to only adopt a single reward signal to represent them. Consequently, there is a growing need for multi-objective reinforcement learning (MORL) (Gábor et al., 1998; Van Moffaert & Nowé, 2014; Yang et al., 2019), which aims to optimize multiple rewards simultaneously.

In MORL, due to the potentially conflicting nature of the objectives, it is generally impossible to find a single policy to maximize all reward functions simultaneously. Therefore, one typically aims to find an optimal policy in the *Pareto* sense, i.e., the performance of any single objective cannot be further improved without compromising other objectives. Also, as a subfield of reinforcement learning (RL), MORL problems can potentially be tackled by various fundamental RL approaches. Among them, the actor-critic approach (Sutton et al., 1998; Konda & Tsitsiklis, 1999) has been widely studied, since it combines the strengths of both value-based and policy-based RL approaches.

When adopting the actor-critic framework for MORL, an immediate challenge stems from how to handle the vector-valued multi-objective reward structure. Toward this end, recent MORL algorithms (e.g., (Nguyen et al., 2020; Chen et al., 2021; Qiu et al., 2024; Zhou et al., 2024b; Ehrgott, 2005;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Fliege et al., 2019)) can be categorized into two major classes: (1) *scalarization methods* (e.g., linear scalarization (LS) and weighted-Chebyshev (WC)) that convert an MORL problem into a single-objective RL problem; and (2) *adaptive gradient methods* (e.g., MGDA method (Désidéri, 2012)) that aim to find a common improving direction.

Despite the aforementioned advances in MORL algorithmic design, the theoretical foundations of MORL remain underdeveloped. The major limitations of the existing MORL theories are **twofold**:

(1) Most of the empirical actor-critic MORL methods incorporate deep neural networks (DNNs) (Yang et al., 2019; Nguyen et al., 2020; Chen et al., 2021). However, existing theoretical analysis techniques for MORL, particularly those concerning convergence and sample complexity (Qiu et al., 2024; Zhou et al., 2024b; Wang et al., 2024), are largely limited to simple linear function approximations or tabular settings. As a result, these analyses are inapplicable to DNN-based MORL actor–critic methods, leading to a significant gap between theory and practice.

(2) Existing theoretical analysis of actor-critic MORL only guarantee convergence to an arbitrary Pareto-stationary policy under linear function approximation (Zhou et al., 2024b; Hairi et al., 2025), without any mechanism for preference control and finite-time global Pareto optimality convergence guarantees. To date, the fundamental problem of preference-guided global convergence to a weak Pareto-optimal policy remains unresolved in actor-critic MORL, let alone the more challenging problem of systematic Pareto front exploration.

The limitations outlined above underscore a substantial gap between the empirical success of DNN-based actor–critic MORL methods and a theoretical foundation for these algorithms. This naturally raises the following question:

(Q): Can we develop actor-critic MORL methods with DNN-based function approximation, which theoretically guarantee 1) achieving global weak Pareto optimality convergence and 2) preference-guided Pareto front exploration?

2) Technical Challenges. Answering the above question is highly non-trivial and necessitates addressing the following technical challenges arising in actor-critic MORL:

- To achieve finite-time convergence rate guarantees, most existing theoretical works on actor–critic MORL (Zhou et al., 2024b; Hairi et al., 2025) adopt an MGDA-type approach from multi-objective optimization (MOO) to dynamically combine the policy gradients of all objectives. This algorithmic choice is motivated by the fact that MGDA naturally generalizes gradient descent from single-objective optimization, enabling tractable convergence analysis. However, MGDA-type methods only guarantee convergence to Pareto-stationary policies, making them fundamentally *unsuitable* for systematic exploration of the Pareto front, which requires weak Pareto-optimal policies. This gap calls for fundamentally new approaches that achieve **global convergence** to weak Pareto optimality, while still retaining provable finite-time convergence guarantees.
- Although several works have established global convergence guarantees for actor–critic single-objective RL under linear function approximation, extending these results to MORL with DNN-based nonlinear function approximation remains an open problem. The complex computational structure introduced by deep neural networks fundamentally renders the proof and analysis techniques developed for the linear setting *inapplicable.* As will be shown later, this difficulty stems from the failure of the Fisher non-degeneracy (FND) condition (a key requirement for proving global convergence under linear function approximation) to hold in DNN-based nonlinear settings. Without the FND condition, the convergence error bounds derived for linear models become *vacuous* in the nonlinear regime.
- Even if finite-time global convergence to weak Pareto optimality can be guaranteed for actor–critic MORL, it remains highly challenging to enable systematic Pareto front exploration guided by preference information. In the MOO literature, the weighted Chebyshev (WC)-scalarization is known for facilitating systematic preference-guided exploration of the Pareto front. However, the inherent *non-smooth* min–max structure of WC-scalarization precludes the existence of well-defined policy gradients. Consequently, WC-scalarization *cannot* be directly integrated into actor-critic MORL, necessitating new methodological approaches for systematic Pareto front exploration for actor-critic MORL.

3) Main Contributions. To overcome the aforementioned challenges, we propose a smoothed weighted-Chebyshev approach (SWEEP) for DNN-based actor-critic MORL, which not only guarantees global convergence to a weak Pareto optimal policy with finite-time convergence rate, but also systematically explores the entire Pareto front. We summarize our contributions as follows:

(1) We show that the two central goals in our DNN-based actor–critic MORL, namely, i) achieving global convergence to weak Pareto-optimal policies and ii) enabling systematic and preference-guided Pareto front exploration, can be elegantly unified within a *new* smoothed weighted Chebyshev (WC) scalarization framework. Specifically, we propose a new smooth approximation of the WC-scalarization that converts the vector-valued MORL problem into a scalar-valued RL problem. This construction resolves the non-differentiability induced by the min–max structure of WC-scalarization, yielding well-defined policy gradients while preserving a one-to-one correspondence between the smoothed WC solutions and the Pareto front, thereby enabling systematic Pareto front exploration.

(2) For the smoothed WC-scalarized single-objective RL problem, we develop a DNN-based nonlinear actor–critic algorithm termed SWEEP, which enjoys a finite-time convergence rate of $O(1/T)$ to a globally optimal policy, where T denotes the total number of iterations. Moreover, leveraging the one-to-one correspondence between the solution set of the smoothed WC-scalarized RL problem and the Pareto front of the original MORL problem, SWEEP further facilitates preference-guided exploration of the Pareto front. To our knowledge, these theoretical guarantees are the first of their kind in the MORL literature.

(3) It is worth noting that, to establish a theoretical finite-time global convergence rate for our SWEEP algorithm, we introduce a series of new assumptions and analytical techniques for studying the convergence behavior of actor–critic reinforcement learning methods. To our knowledge, this line of analysis is novel in the literature and may be of independent interest in RL. Furthermore, to empirically validate the effectiveness of SWEEP, we conduct extensive numerical experiments on both recommendation system training and multi-objective robotic simulation, which demonstrate that SWEEP can efficiently explore the Pareto front.

2 Related Work

In this section, we summarize the related works in MORL and two closely related fields: single-objective actor-critic algorithms, and multi-objective optimization algorithms.

1) Single-Objective Actor-Critic RL Algorithms: The actor-critic framework, along with their variants have been one of the most widely used approaches in RL (Konda & Tsitsiklis, 1999; Peters & Schaal, 2008; Mnih et al., 2016). Besides their empirical successes, several works have also established rigorous theoretical finite-time convergence rate and sample complexity results (Xu et al., 2020; Qiu et al., 2021; Cayci et al., 2024; Tan et al., 2025). Moreover, recent works have begun to explore the DNN-based actor-critic algorithms (Wang & Hu, 2021; Gaur et al., 2024; Zhang et al., 2025b; Ganesh et al., 2025). However, the theories of DNN-based actor-critic approaches for MORL remain largely missing in the literature.

2) Multi-Objective Optimization Algorithms: The history of Multi-Objective Optimization (MOO) problems dates back to (Sawaragi et al., 1985), and recent years have seen increasing development of MOO theories (Ehrgott, 2005; Gunantara, 2018; Sharma & Kumar, 2022). For example, the theoretical understanding for MOO approaches such as weighted-Chebyshev (WC, Momma et al. (2022); Lin et al. (2024); Zhang et al. (2025a)), multi-gradient descent algorithm (MGDA, Désidéri (2012); Xu et al. (2025)) have been established. However, when being applied in DNN-based actor-critic MORL, the theoretical convergence results of these MOO approaches remain unclear.

3) MORL Algorithms: Compared to the previous two areas, the theoretical studies on MORL only started in recent years. Although several types of MORL algorithms have been investigated (e.g., (Yang et al., 2019; Zhou et al., 2024b; Qiu et al., 2024; Wang et al., 2024; Hairi et al., 2025)), their theoretical convergence results remain poorly understood. In particular, most of the existing works failed to address at least one of the following aspects: 1) the use of DNN-based nonlinear function approximations, 2) providing finite-time global convergence guarantees, and 3) exploring the entire weak Pareto optimal front. In contrast, our algorithm significantly advances MORL by simultaneously addressing all of these technical barriers. Due to space limitation, we relegate a detailed comparison and more related works in the literature of MORL to Appendix A.

3 MULTI-OBJECTIVE REINFORCEMENT LEARNING: A PRIMER

In this section, we will begin by formally formulating the MORL problem and providing several key definitions in MORL. Next, we will introduce the DNN-based actor-critic approach for MORL. Finally, we propose a new smooth approximation of the WC-scalarized problem for solving DNN-based actor-critic MORL.

3.1 THE MORL PROBLEM

Consider an m-objective Markov decision process (MOMDP): $(\mathcal{S}, \mathcal{A}, \{r_i\}_{i=1}^m, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\{r_i\}_{i=1}^m$ is the reward signal vector with $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for the i-th objective, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel, and $\gamma \in (0, 1)$ is the discount factor. Here, we adopt the so-called "restart transition kernel" that has been widely used in the literature (e.g., (Xu et al., 2020; Chen et al., 2022)). Specifically, $\mathcal{P}$ is defined as $\mathcal{P}(s, a, s') := \gamma\mathbb{P}(s'|s, a) + (1 - \gamma)\mathbb{I}\{s' = s_0\}$, where s_0 denotes the initial state. An MORL policy is denoted by $\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0, 1]$, where θ represents its parameters. Hence, for each objective $i \in [m]$, we define the cumulative discounted reward for policy π_θ as $J_i(\theta) := \mathbb{E}_{\pi_\theta}\big[\sum_{t=0}^{\infty} \gamma^t r_{i,t}\big]$. The MORL problem can thus be formulated as:

$$\max_\theta \mathbf{J}(\theta) = \big[J_1(\theta), \ldots, J_m(\theta)\big]^\top. \tag{1}$$

In MORL, the vector-valued objective is usually associated with a preference weight vector $p \in \Delta_m^+$, where $\Delta_m^+ := \{p \in \mathbb{R}^m : p \geq 0, \sum_{i=1}^m p_i = 1\}$ denotes the standard m-dimensional simplex, which represents potentially different attention on each objective. As mentioned earlier, since it is impossible to maximize multiple objectives with a single policy π_θ in general, we introduce the following optimality criterion for solving MORL problems:

Definition 1 (Pareto Optimality)**.** Policy π_θ dominates $\pi_{\theta'}$ if and only if $J_i(\theta) \geq J_i(\theta'), \forall i \in [m]$, and $J_i(\theta) > J_i(\theta'), \exists i \in [m]$. Policy π_θ is Pareto optimal if no other policy $\pi_{\theta'}$ dominates π_θ. Also, policy π_θ is weakly Pareto optimal if no other policy $\pi_{\theta'}$ satisfies: $J_i(\theta') > J_i(\theta), \forall i \in [m]$.

Clearly, Pareto optimality implies weak Pareto optimality, while the converse is not true. We can interpret Pareto optimality as the inability to find a policy that improves the performance of each objective simultaneously. Moreover, we also denote the set of all Pareto optimal (resp. weakly Pareto optimal) policies as Θ_{P} (resp. Θ_{WP}), and the Pareto front (resp. weak Pareto front) as $\{F(\theta) : \theta \in \Theta_{\text{P}}\}$ (resp. $\{F(\theta) : \theta \in \Theta_{\text{WP}}\}$).

3.2 THE DEEP-NEURAL ACTOR-CRITIC APPROACH FOR MORL

Next, we will introduce the basics of the actor-critic MORL, which is followed by the deep-neural actor-critic approach for MORL.

1) The Actor-Critic MORL Framework: The actor-critic MORL framework involves two stages: First, for some given policy, the critic component evaluates its value function, indicating the "goodness" of that policy; Second, based on the approximated value function, the actor component updates the policy using policy gradients. Specifically, the value function is defined as:

$$V_{\theta,i}(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r_{i,t} \Big| s_0 = s, a_t \sim \pi_\theta(\cdot|s_t)\Big], \quad i \in [m].$$

However, the true value of $V_{\theta,i}(s)$ is unknown in general during the MORL process. Thus, the critic component approximates the state value function as $\widehat{V}(s; W_i)$ using techniques such as TD-learning, where W_i denotes the parameters of the critic model. In this paper, W_i is assumed to be parameterized by DNNs, which will be discussed later.

We define the advantage function $\text{Adv}_{\theta,i}(s, a) := r(s, a) + \gamma\mathbb{E}\big[V_{\theta,i}(s')\big|s' \sim \mathcal{P}(\cdot|s, a)\big] - V_{\theta,i}(s)$. The actor can compute the policy gradients by utilizing the following policy gradient theorem (Xu et al., 2020; Zhang et al., 2025b):

Lemma 1 (Policy Gradient Theorem)**.** *Under the restart kernel $\mathcal{P}$, for any policy π_θ and for any $i \in [m]$, the policy gradient satisfies: $\nabla_\theta J_i(\theta) \propto \mathbb{E}\big[\nabla_\theta \log \pi_\theta(a|s) Adv_{\theta,i}(s, a)\big|(s, a) \sim \nu(\theta)\big]$, where $\nu(\theta)$ denotes the stationary distribution, for which the existence will be specified in Assumption 2.*

2) Deep-Neural Actor-Critic Methods for MORL: As mentioned earlier, no theoretical foundations have been established for DNN-based actor-critic MORL in the literature. However, there do exist some theoretical foundations for DNN-based actor-critic approaches for single-objective RL (Cai et al., 2019; Gaur et al., 2024; Zhang et al., 2025b). Similar to these existing works, we also adopt multi-layer perceptron (MLP) architecture for DNN-based actor-critic for MORL in this paper. Specifically, we encode each state s by some $x \in \mathbb{R}^d$ with a one-to-one mapping, where it is assumed that $\|x\|_2 = 1$ without loss of generality. Then, a DNN can be represented as follows:

$$x^{(0)} = Ax, \qquad y = c^\top x^{(D)}, \qquad x^{(b)} = \frac{1}{\sqrt{w}}\mathsf{Activation}(W^{(b)}x^{(b-1)}), \forall b \in [D],$$

where $A \in \mathbb{R}^{w\times d}$, $W^{(b)} \in \mathbb{R}^{w\times w}, \forall b \in [D]$, and $c \in \mathbb{R}^w$ are the parameters of the DNN, and $\mathsf{Activation}(\mathbf{v})$ denotes the activation function. In this paper, we adopt the Sigmoid activation, which is defined as $\mathsf{Sigmoid}(\mathbf{v}) := \frac{\mathbf{1}}{\mathbf{1}+\exp(-\mathbf{v})}$. According to (Shen et al., 2022; Zhang et al., 2024), we have the following universal approximation result for the Sigmoid-DNN with depth D and width w, which will be useful in our theoretical analysis:

Lemma 2 (Universal Approximation)**.** *Suppose $V_{\theta,i}(x)$ is Lipschitz continuous (see Assumption 3). Then, there exists a* Sigmoid*-DNN parameterized by W, such that: $\max_{x,i,\theta}\left|\widehat{V}(x;W) - V_{\theta,i}(x)\right| = \widetilde{\mathcal{O}}\big(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\big)$, where $\widetilde{\mathcal{O}}(\cdot)$ hides the constants and logarithmic terms.*

Lemma 2 says that, when $V_{\theta,i}$ is Lipschitz continuous, as the width and depth of the DNN increase, the approximation error vanishes at nearly a rate of $1/(wD)^{\frac{2}{d}}$. It is also worth pointing out that our choice of the Sigmoid activation does *not* lose any generality in our subsequent theoretical analysis because universal approximation properties with the same approximation rate also hold for MLPs with activation functions (e.g., Tanh and ReLU) (Shen et al., 2022; Zhang et al., 2024). We note that, after initializing all entries of A and $W^{(b)}, \forall b \in [D]$ independently following $\mathcal{N}(0,2)$, and those in c independently following $\mathcal{N}(0,1)$, we only update $W = \big(W^{(1)}, \ldots, W^{(D)}\big)$ during training. We thus simplify the notation $\widehat{V}(x;W,A,c)$ to $\widehat{V}(x;W)$.

3.3 Weighted-Chebyshev Scalarization for MORL

Lastly, we conclude this section by introducing the weighted-Chebyshev (WC)-scalarization technique, which transforms a vector-valued multi-objective optimization problem (MOO) into a conventional scalar-valued optimization problem. This facilitates the use of existing optimization tools for solving MOO problems. The most salient feature of WC-scalarization is that, by associating and varying a preference vector in the m-dimensional standard simplex, one can systematically explore the entire weak Pareto front of the MOO problem.

To formally derive the WC-scalarization for MORL, we first transform the "reward maximization" in MORL in Eq. (1) into a "regret minimization" form to be conformal to the polarity of the standard WC-scalarization. Toward this end, let J_i^{ub} denote an upper bound of $J_i(\theta)$. Then, we can reformulate the MORL problem as follows[1]: $\min_\theta\{\mathbf{J}^{\text{ub}} - \mathbf{J}(\theta)\} = \big[J_1^{\text{ub}} - J_1(\theta), \ldots, J_m^{\text{ub}} - J_m(\theta)\big]^\top$, where $J_i^{\text{ub}} - J_i(\theta) > 0, \forall i \in [m], \theta$. For a preference vector $p \in \Delta_m^+$, the WC-scalarized problem for MORL is defined as follows:

$$\min_\theta \|p \odot (\mathbf{J}^{\text{ub}} - \mathbf{J}(\theta))\|_\infty = \min_\theta\{\max_i\{p_i(J_i^{\text{ub}} - J_i(\theta))\}\}.$$

Although WC-scalarization facilitates Pareto front exploration, its "min-max" structure induces non-smoothness and complicates algorithmic design.

4 The SWEEP Algorithm

1) Algorithm Overview: Our DNN-based smoothed weighted-Chebyshev approach (SWEEP), presented in Algorithm 1, solves the smoothed WC-scalarized problem with a double-loop structure.

[1]This transformation does not lose any generality, as the Pareto fronts of these two problems have a 1-to-1 correspondence. Also, the 1-to-1 correspondence implies that the choice of $\mathbf{J}^{\text{ub}}$ is flexible as long as it is a proper upper bound. Thus, one may choose a trivial upper bound that is easy to work with, such as $r_{\max}/(1-\gamma)$ if all rewards are upper bounded by $r_{\max}$.

Algorithm 1 SWEEP: DNN-based Actor-Critic for MORL.

Input: step-size α_t, initial parameters θ_0, initial state s_0, preference p, $\mathbf{J}^{\text{ub}}$, and μ.
for $t = 0, 1, \ldots, T-1$ **do**
  Let $s_{t_0}, \{W_{i,t}\}_{i=1}^m$ be output of Alg. 2.
  for $l = 0, 1, \ldots, M-1$ **do**
    Observe $s_{t_{l+1}}$ and $r_{i,t_{l+1}}, i \in [m]$.
    Sample $a_{t_{l+1}} \sim \pi_{\theta_t}(\cdot|s_{t_{l+1}})$.
    Compute $\psi_{t_l} = \nabla_\theta \log \pi_{\theta_t}(s_{t_l}, a_{t_l})$.
    for $i \in [m]$ **do**
      Compute $\delta_{i,t_l} = \widehat{V}(s_{t_l}; W_{i,t}) - r_{i,t_{l+1}} - \gamma \widehat{V}(s_{t_{l+1}}; W_{i,t})$.
    end for
  end for
  for $i \in [m]$ **do**
    Compute $\widehat{\nabla} J_i(\theta_t) = \frac{1}{M} \sum_{l=0}^{M-1} \delta_{i,t_l} \psi_{t_l}$, and $\widehat{J}_i(\theta_t) = \widehat{V}(s_0; W_{i,t})$.
  end for
  Compute d_t according to Equation (3).
  Update $\theta_{t+1} = \theta_t - \alpha_t \frac{d_t}{\|d_t\|}$.
end for
Output: Policy θ_T.

Algorithm 2 DNN-Based Critic for MORL.

Input: s_0, π_{θ_t}, step-size β, iteration steps K, projection radius B.
Initialize: $\mathcal{B}(B) = \{W : \|W^{(b)} - W^{(b)}(0)\|_{\text{F}} \le B, \forall b \in [D]\}$. $W_i(0) = W_i = W(0), \forall i \in [m]$.
for $k = 0, 1, \ldots, K-1$ **do**
  Sample $(s_k, a_k, \{r_{i,k+1}\}_{i=1}^m, s_{k+1}, a_{k+1})$, where $a_k \sim \pi_{\theta_t}(\cdot|s_k)$.
  for $i \in [m]$ **do**
    Compute TD-error $\delta_{i,k} = \widehat{V}(s_k; W_i(k)) - r_{i,k+1} - \gamma \widehat{V}(s_{k+1}; W_i(k))$.
    Update $\widetilde{W}_i(k+1) = W_i(k) - \beta \delta_{i,k} \cdot \nabla_W \widehat{V}(s_k; W_i(k))$.
    Project $W_i(k+1) = \arg\min_{W \in \mathcal{B}(B)} \|W - \widetilde{W}_i(k+1)\|_2$.
    Update $W_i = \frac{k+1}{k+2} W_i + \frac{1}{k+2} W_i(k+1)$.
  end for
end for
Output: $s_{K-1}, \{W_i\}_{i=1}^m$.

In the inner-loop, the critic component iterates for K steps, leveraging TD-learning method to approximate the m value functions $V_{\theta,i}$ for the current policy θ with m DNNs. The outer-loop executes T rounds in total, where in each round, the actor component approximates the gradient of G_μ using the obtained value functions, and then updates the policy θ, which is also parameterized by a DNN with the same width and depth.

2) The Critic Component: The critic component is presented in Algorithm 2, which aims to compute a value function approximation $\widehat{V}(x; W_i)$ for each objective $i \in [m]$. Specifically, for the current policy θ_t and the i-th objective, the critic first computes the TD-error $\delta_{i,k}$ at step k, then performs a TD update. The newly obtained parameters are then projected onto a ball centered at $W(0)$ with radius B, i.e., $\mathcal{B}(B)$. This projection ensures the non-expansive property of the convex ball and is useful in the subsequent analysis.

3) The Actor Component: Each actor step t begins with a Markov batch sampling with a batch-size of M. During this process, the algorithm maintains the score function ψ_{t_l} and TD-error δ_{i,t_l} for each $i \in [m]$. Upon the completion of Markov batch sampling, we approximate each objective function $J_i(\theta_t)$ and its gradient $\nabla J_i(\theta_t)$. However, note that the WC-scalarization is in the "min-max" form, which is *non-smooth* and results in ill-defined gradient for the WC objective. To address this challenge, we adopt a smoothed WC-scalarization defined as follows (Lin et al., 2024):

$$\min_\theta G_\mu(\theta|p) := \mu \log \left(\sum_{i=1}^m \exp \frac{p_i(J_i^{\text{ub}} - J_i(\theta))}{\mu} \right), \tag{2}$$

where $\mu > 0$ is a tunable hyperparameter. It is shown in (Lin et al., 2024) that the smoothed WC approximation can approximate the original WC-scalarization and maintain desirable properties:

Lemma 3 (Pareto Front Reconstruction)**.** *There exists some constant $\mu_0 > 0$, such that, for any fixed $\mu \in (0, \mu_0]$, the policy θ is a weakly Pareto optimal policy if and only if it is the solution of Eq. (2) under some preference $p \in \Delta_m^{++}$, where Δ_m^{++} is the positive standard m-simplex.*

This property highlights that solving the smoothed WC approximation problem to optimality and varying p across the m-simplex still allow us to explore the entire weak Pareto front. Let $\theta^*(p, \mu)$ be the minimizer of Eq. (2). Solving the MORL problem is then equivalent to solving this scalar-valued

and smoothed WC-scalarized problem, i.e., finding $\theta^*(p,\mu)$ for a $p \in \Delta_m^{++}$. Let $\widehat{J}_i(\theta_t)$ and $\widehat{\nabla} J_i(\theta_t)$ denote the approximations of i-th state-value function and the policy gradient obtained by using Lemma 1, respectively. Then, it can be shown that one can leverage these results to estimate the policy gradient $\nabla G_\mu(\theta_t|p)$ by substituting the ground truth with our approximations as follows:

$$d_t = -\sum_{i=1}^{m} \frac{\exp(\frac{p_i(J_i^{\text{ub}} - \widehat{J}_i(\theta_t))}{\mu})}{\sum_{i'=1}^{m} \exp(\frac{p_{i'}(J_{i'}^{\text{ub}} - \widehat{J}_{i'}(\theta_t))}{\mu})} p_i \widehat{\nabla} J_i(\theta_t). \tag{3}$$

Then, the policy θ_{t+1} is updated in a gradient descent manner using $G_\mu(\theta_t|p)$. Three important remarks are in order.

Remark 1. While our SWEEP algorithm follows the actor-critic framework, the key **novelty** and **difference** lie in the policy update step. Specifically, after using policy gradient theorem (i.e., Lemma 1) to approximate $\widehat{\nabla} J_i(\theta_t)$, SWEEP does not directly perform a gradient descent step on $J_i(\theta_t)$. Instead, we utilize $J_i(\theta_t)$ to further approximate $\nabla G_\mu(\theta_t \mid p)$ and then perform a gradient update. These structural differences lead to drastic different proofs and analysis.

Remark 2. We note that in several existing MORL works, the actor component utilizes the MGDA technique (Zhou et al., 2024b; Wang et al., 2024; Hairi et al., 2025). In particular, after approximating $\nabla J_i(\theta_t)$ by $\widehat{\nabla} J_i(\theta_t)$, these works solve a quadratic programming $\min_\lambda \|\lambda \odot \widehat{\nabla} J_i(\theta_t)\|_2^2$ to determine a common descent direction. While these MGDA-based approaches enable finite-time convergence rate analysis (Désidéri, 2012), the inherent limitations of MGDA only guarantee a convergence to a Pareto stationary policy, rather than a global convergence to a weak Pareto optimal policy. In contrast, through the smoothed WC-scalarization problem in Eq. (2), SWEEP is able to exploit the special properties of the policy gradients combined with Lemma 3, which play a key role in ensuring global convergence to weak Pareto optimal policies and systematic Pareto front exploration.

Remark 3. We note that most existing theoretical works on actor-critic RL literature use the Q-function (Cai et al., 2019; Gaur et al., 2024; Zhang et al., 2025b). A key **difference** in our SWEEP method is that it employs V-function approximation (Chen et al., 2022; Hairi et al., 2022; Zhou et al., 2024b). This design offers significant benefits in practical MORL implementation. To see this, note that in Line 13 of Algorithm 1, we need to approximate $\widehat{J}_i(\theta_t)$ for each objective i (unique to the smoothed WC-scalarized MORL approach and unseen in the previous literature). Using the Q-approximation $\widehat{J}_i(\theta_t) = \sum_a \widehat{Q}(s_0, a; W_{i,t})\pi_{\theta_t}(a|s_0)$ requires enumerating the entire action space, which is impractical or even infeasible in MORL problems with large or continuous action space (e.g., video streaming recommendation systems and LLM alignment). In contrast, our V-approximation circumvents this difficulty and substantially enhances the efficiency in solving the MORL problem.

5 Theoretical Convergence Analysis

In this section, we analyze SWEEP's convergence performance. We begin by stating some useful assumptions, which will be followed by our main theoretical results on finite-time global convergence and their insights.

Assumption 1 (Reward)**.** There exists some $r_{\max} > 0$ such that, $r_{i,t} \in [0, r_{\max}], \forall t \geq 0, i \in [m]$.

Assumption 2 (Geometric Mixing Time)**.** For any policy π_θ, there exists a stationary distribution $\nu(\theta)$ for (s,a). Further, for any policy π_θ, there exist positive constants $\eta, \rho \in (0,1)$ so that $\sup_{s\in\mathcal{S}} \|\mathbb{P}\big(s_t, a_t \big| s_0 = s\big) - \nu(\theta)\|_{\text{TV}} \leq \eta\rho^t, \forall t \geq 0$.

Assumption 2 is standard in the literature (Zou et al., 2019b; Xu et al., 2020; Gaur et al., 2024; Wang et al., 2024). Notably, the stationary distribution and the mixing behavior of the MOMDP can also be equivalently ensured by assuming the irreducible and aperiodic MOMDP (Hairi et al., 2022; Zhou et al., 2024b; Zhang et al., 2025b).

Assumption 3 (Lipschitz Continuity)**.** $J_i(\theta)$ and $\nabla J_i(\theta)$ are Lipschitz continuous, i.e., there exist two positive constants L_J and M_J such that, for any $i \in [m]$, and for any θ and θ', we have $|J_i(\theta) - J_i(\theta')| \leq L_J\|\theta - \theta'\|_2$, $\|\nabla J_i(\theta) - \nabla J_i(\theta')\|_2 \leq M_J\|\theta - \theta'\|_2$. Additionally, $V_{\theta,i}(x)$ is L_V-Lipschitz continuous, i.e., for any x, x', i and θ, we have $|V_{\theta,i}(x) - V_{\theta,i}(x')| \leq L_V\|x - x'\|_2$.

Assumption 3 is also commonly adopted in the literature (Wang et al., 2024; Zhou et al., 2024b; Gaur et al., 2024; Zhang et al., 2025b), which i) allows us to apply the descent lemma in theoretical convergence analysis, and ii) ensures the universal approximation result presented in Lemma 2.

Assumption 4. For any s, a and any policy θ, there exist a positive constant M_g such that the score function satisfies: $\|\nabla_\theta \log \pi_\theta(a|s)\|_2 \leq M_g$.

Assumption 5. For any policy θ and for $i \in [m]$, there exist positive constants σ and ϵ_{bias} such that $\mathbb{E}\big(\text{Adv}_{\theta,i}(s,a) - (1-\gamma)(F(\theta)^{-1}\nabla_\theta J_i(\theta))^\top \nabla_\theta \log \pi_\theta(a|s)\big)^2 \leq \epsilon_{\text{bias}}$, where $(s,a) \sim \nu(\pi_i^*)$ (stationary distribution under the optimal policy π_i^* with respect to objective i), $F(\theta) = \mathbb{E}_{\nu(\theta)}(\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top) + \sigma I$.

We note that Assumption 5 is **new** and relaxes the Fisher non-degeneracy (FND) condition (i.e., $\mathbb{E}_{\nu(\theta)}(\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top) > \sigma$), which has been widely used in the literature on actor-critic RL with linear function approximation to obtain global convergence (see, e.g., (Liu et al., 2020; Ding et al., 2022; Masiha et al., 2022)). The fundamental rationale behind proposing Assumption 5 is that, in actor-critic with MLP-based function approximation, the FND assumption does *not* hold since the null space of $\nabla_\theta \log \pi_\theta(a|s)$ can be non-empty, preventing the Fisher information matrix from full rank. To overcome this challenge, we remove the conventional FND assumption and instead introduce a σ-perturbation in our Assumption 5, which enables us to establish global convergence for actor-critic MORL with MLP-based function approximation. We note that Assumption 5 could be *of independent interest* in the literature on actor-critic RL with nonlinear function approximation.

Assumption 6. For $p \in \Delta_m^{++}$, $\mu > 0$, there exists some $\rho > 0$ such that, for any θ, we have: $\frac{1}{2}\|\nabla G_\mu(\theta \mid p)\|^2 \geq \rho[G_\mu(\theta \mid p) - G_\mu(\theta^*(p,\mu) \mid p) - \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}]_+$.

Assumption 6 is a relaxed PL-style condition, which is widely adopted in the literature. With these assumptions, we are now ready to state our main results.

Theorem 1 (Global Convergence)**.** *Under Assumptions 1–6, with sufficiently small μ, $\alpha_t = \frac{\alpha}{t}$, $\alpha \geq \max\{1, \frac{M_g L_J \sqrt{m}}{\sigma \delta_0} + \frac{\mu \log m \sqrt{m}}{\delta_0}\}$, $\beta = 1/\sqrt{K}$, $w = \Omega(d^3 D^{-\frac{11}{2}})$, $B = \Theta(w^{\frac{1}{32}} D^{-6})$, and $K = \Omega(D^4)$, Algorithm 1 achieves the following global convergence guarantee with probability at least $1 - \exp^{-\Omega(\log^2 w)}$ for any $p \in \Delta_m^{++}$:*

$$\mathbb{E}\Big[G_\mu(\theta_T \mid p) - G_\mu(\theta^*(p,\mu) \mid p)\Big] = \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\sqrt{\epsilon_{bias}}\right) + \mathcal{O}\left(M^{-\frac{1}{2}}\right)$$
$$+ \mathcal{O}\left(w^{-\frac{2}{d}} D^{-\frac{2}{d}} \mu^{-1} m^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(w^{\frac{1}{32}} D^{-\frac{7}{2}} K^{-\frac{1}{4}} \mu^{-1} m^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{1}{24}} D^{-4} \mu^{-1} m^{\frac{1}{2}}\right).$$

Our key idea in proving global convergence in Theorem 1 is to establish a counterpart of the *performance difference lemma* for the single-objective RL (Kakade & Langford, 2002) in our MORL setting (cf. Appendix B). This property is based on a well-defined gradient of the objective function, which unfortunately does not hold under WC-scalarization. By utilizing the smoothed approximation of WC-scalarization, we are able to derive a smooth scalar-valued objective, which yields the performance difference lemma tailored for MORL.

Theorem 1 says that Algorithm 1 converges to a global minimum of the smoothed WC-scalarized solution of Eq. (2) at a convergence rate of $\mathcal{O}(1/T)$. Moreover, by Lemma 3, achieving this smoothed WC-scalarized global minimum implies obtaining a weakly Pareto optimal policy for the original MORL problem. Furthermore, Algorithm 1 also explores the entire Pareto front Θ_{WP} by varying the preference vector p in the m-simplex Δ_m^{++}. To our knowledge, these finite-time global convergence and Pareto front exploration results are the *first of their kind* in the MORL literature.

Additionally, Theorem 1 provides insights into how DNNs affect performance. It can be seen from the convergence error bound in Theorem 1 that increasing the depth D of the DNNs significantly improves the performance of Algorithm 1, whereas increasing the width w has a negligible impact.

6 Numerical Experiments

In this section, we conduct numerical experiments to verify the performance of our SWEEP method using multi-objective robotic simulations and recommendation systems.

1) Experiments on Robotic Simulation: *1-1) Experimental Setup:* For the ease of validating SWEEP's Pareto front exploration, we use a robotic simulation task under the MuJoCo-Walker-2d-v5 environment (Felten et al., 2023) with a bi-objective problem, where the *Move* objective encourages the walker to move forward, while the *Control* objective aims to minimize the control

Table 1: Comparison of our method with baseline methods.

Algorithm	Click↑	Like↑(e-2)	Follow↑(e-4)	Comment↑(e-3)	Forward↑(e-3)	Dislike↓(e-4)	WatchTime↑
Behavior-Clone	0.534	1.231	4.608	3.225	**1.119**	2.304	1.285
MOAC (Zhou et al., 2024b)	0.535	1.261	4.946	2.780	1.105	1.395	1.249
(Linear approx.)	0.30%	2.46%	7.33%	−13.8%	−1.23%	−39.4%	−2.84%
MOCHA (Hairi et al., 2025)	0.535	1.348	4.109	3.033	1.020	1.373	1.235
(Linear approx.)	0.15%	9.48%	−10.8%	−5.97%	−8.86%	−40.4%	−3.94%
PDPG (Chen et al., 2021)	**0.539**	1.228	4.828	3.165	0.919	**1.140**	**1.308**
(DNN)	**1.02%**	−0.26%	4.78%	−1.86%	−17.8%	**−50.5%**	**1.74%**
SWEEP (Ours)	**0.539**	**1.372**	**5.042**	**3.324**	0.960	1.538	1.293
(DNN)	**1.02%**	**11.47%**	**9.42%**	**3.07%**	−14.19%	−33.24%	0.58%

effort. To validate the systematical Pareto exploration capability of our approach, we randomly sample various preference vectors across the standard simplex Δ_2^+. In addition, we also conduct ablation studies within this Walker environment to demonstrate how our algorithm outperforms the linear scalarization-based actor-critic method.

1-2) Experimental Results: In Fig. 1, each point represents the rewards of two objectives under a fixed preference vector p. The blue and yellow curves are obtained by our SWEEP and the linear scalarization-based actor-critic methods, respectively. Clearly, SWEEP dominates the linear scalarization actor-critic approach, highlighting the effectiveness of the smoothed-WC technique in Pareto front exploration. In addition to Fig. 1, we further evaluate the performance of SWEEP using the hypervolume metric. The hypervolume values achieved by SWEEP and the linear-scalarization-based actor–critic method are 2.936 and 2.213, respectively, indicating superior Pareto exploration performance of SWEEP. Due to space limitations, we relegate the detailed experiment setups and additional results on the hypervolume metrics and the impacts of DNN's sizes to Appendix C.2.

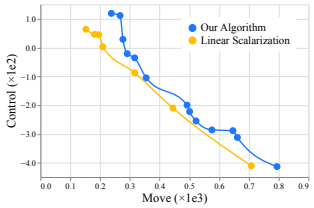


Figure 1: Visualization of Pareto fronts achieved by our algorithm and linear scalarization-based actor-critic method.

2) Experiments on Recommendation Systems: *2-1) Experimental Setup:* We next conduct larger-scale experiments on the Kuairand dataset (Gao et al., 2022), an unbiased sequential dataset collected from the recommendation logs of a video-sharing app, which has seven potentially conflicting rewards: Click, Like, Follow, Comment, Forward, Dislike, and Watch Time. We compare our approach with three state-of-the-art theoretical MORL baselines: PDPG, MOAC, and MOCHA.

2-2) Experimental Results: Our comparison results are summarized in Table 1, where Behavior-Clone, a supervised method that mimics real customer behavior, serves as the benchmark for all MORL methods. We observe that: (i) the utilization of DNN significantly enhances overall performance, highlighting the strength of DNN over linear approximation; (ii) our SWEEP method outperforms the DNN-based PDPG on metric *Like*, *Follow*, and *Comment*, while PDPG exhibits slightly better results on *Dislike* and *Watch Time*. To evaluate our method on Pareto front boundary discovery, we choose a group of seven unit preference vectors, each exhibiting a maximal preference toward a specific objective. Fig. 2 shows that by varying the preference vectors, SWEEP can explore the boundary of the Pareto front. Due to space limitation, the detailed setups and additional numerical results are relegated to Appendix C.1.

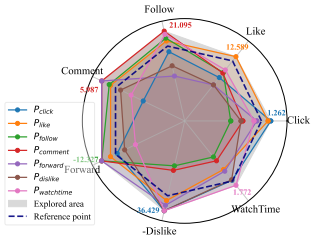


Figure 2: Pareto front boundary discovery.

7 Conclusion

In this paper, we proposed the SWEEP algorithm based on a smoothed weighted-Chebyshev scalarization for DNN-based actor-critic MORL. We showed that SWEEP converges to a global weak Pareto optimal policy at a finite-time convergence rate of $\mathcal{O}(1/T)$, establishing the first theoretical guarantees for solving DNN-based actor-critic MORL problems. Moreover, we showed that SWEEP enables effective Pareto front exploration.

AI USE STATEMENT

We used generative AI tools solely to polish the language. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

We have not identified any specific ethical concerns that require additional discussion in this work.

REPRODUCIBILITY STATEMENT

We may consider making data and code openly accessible when necessary.

REFERENCES

Nihal Acharya Adde, Alexandra Gianzina, Hanno Gottschalk, and Andreas Ebert. Robust evolutionary multi-objective network architecture search for reinforcement learning (emnas-rl). *arXiv preprint arXiv:2506.08533*, 2025.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE signal processing magazine*, 34(6):26–38, 2017.

Hui Bai, Ran Cheng, and Yaochu Jin. Evolutionary reinforcement learning: A survey. *Intelligent Computing*, 2:0025, 2023.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

Semih Cayci, Niao He, and R Srikant. Finite-time analysis of natural actor-critic for pomdps. *SIAM Journal on Mathematics of Data Science*, 6(4):869–896, 2024.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Xu Chen, Yali Du, Long Xia, and Jun Wang. Reinforcement recommendation with user multi-aspect preference. In *Proceedings of the Web Conference 2021*, pp. 425–435, 2021.

Ziyi Chen, Yi Zhou, Rong-Rong Chen, and Shaofeng Zou. Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis. In *International Conference on Machine Learning*, pp. 3794–3834. PMLR, 2022.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 1910–1934. PMLR, 2022.

Matthias Ehrgott. *Multicriteria optimization*. Springer, 2005.

Florian Felten, Lucas N Alegre, Ann Nowe, Ana Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:23671–23700, 2023.

Florian Felten, El-Ghazali Talbi, and Grégoire Danoy. Multi-objective reinforcement learning based on decomposition: A taxonomy and framework. *Journal of Artificial Intelligence Research*, 79: 679–723, 2024.

Jörg Fliege, A Ismael F Vaz, and Luís Nunes Vicente. Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34(5):949–959, 2019.

Zoltán Gábor, Zsolt Kalmár, and Csaba Szepesvári. Multi-criteria reinforcement learning. In *ICML*, volume 98, pp. 197–205, 1998.

Swetha Ganesh, Jiayu Chen, Washim Uddin Mondal, and Vaneet Aggarwal. Order-optimal global convergence for actor-critic with general policy and neural critic parametrization. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624. URL https://doi.org/10.1145/3511808.3557624.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (last iterate) of actor-critic under markovian sampling with neural network parametrization. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=rJxFvAs7pq.

Nyoman Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018.

FNU Hairi, Jia Liu, and Songtao Lu. Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward," in proc. iclr, virtual event, april 2022. *Proc. ICLR*, 2022.

Fnu Hairi, Jiao Yang, Tianchen Zhou, Haibo Yang, Chaosheng Dong, Fan Yang, Michinari Momma, Yan Gao, and Jia Liu. Enabling pareto-stationarity exploration in multi-objective reinforcement learning: A multi-objective weighted-chebyshev actor-critic approach. *arXiv preprint arXiv:2507.21397*, 2025.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *arXiv preprint arXiv:2103.09568*, 2021.

Hannah Janmohamed and Antoine Cully. Multi-objective quality-diversity in unstructured and unbounded spaces. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 149–157, 2025.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Shauharda Khadka, Somdeb Majumdar, Tarek Nassar, Zach Dwiel, Evren Tumer, Santiago Miret, Yinyin Liu, and Kagan Tumer. Collaborative evolutionary reinforcement learning. In *International conference on machine learning*, pp. 3341–3350. PMLR, 2019.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30479–30509, 2024.

Erlong Liu, Yu-Chang Wu, Xiaobin Huang, Chengrui Gao, Ren-Jian Wang, Ke Xue, and Chao Qian. Pareto set learning for multi-objective reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18789–18797, 2025.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Saeed Masiha, Saber Salehkaleybar, Niao He, Negar Kiyavash, and Patrick Thiran. Stochastic second-order methods improve best-known sample complexity of sgd for gradient-dominated functions. *Advances in Neural Information Processing Systems*, 35:10862–10875, 2022.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PmLR, 2016.

Michinari Momma, Chaosheng Dong, and Jia Liu. A multi-objective/multi-task learning framework induced by pareto stationarity. In *International Conference on Machine Learning*, pp. 15895–15907. PMLR, 2022.

David E Moriarty, Alan C Schultz, and John J Grefenstette. Evolutionary algorithms for reinforcement learning. *Journal of Artificial Intelligence Research*, 11:241–276, 1999.

Thanh Thi Nguyen, Ngoc Duy Nguyen, Peter Vamplew, Saeid Nahavandi, Richard Dazeley, and Chee Peng Lim. A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence*, 96:103915, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

Shuang Qiu, Dake Zhang, Rui Yang, Boxiang Lyu, and Tong Zhang. Traversing pareto optimal policies: Provably efficient multi-objective reinforcement learning. *arXiv preprint arXiv:2407.17466*, 2024.

Diederik Roijers, Denis Steckelmacher, and Ann Nowé. Multi-objective reinforcement learning for the expected utility of the return. In *Adaptive Learning Agents Workshop 2018*, 2018.

Yoshikazu Sawaragi, HIROTAKA NAKAYAMA, and TETSUZO TANINO. *Theory of multiobjective optimization*, volume 176. Elsevier, 1985.

Shubhkirti Sharma and Vijay Kumar. A comprehensive review on multi-objective optimization techniques: Past, present and future. *Archives of Computational Methods in Engineering*, 29(7): 5605–5633, 2022.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Tianye Shu, Ke Shang, Cheng Gong, Yang Nan, and Hisao Ishibuchi. Learning pareto set for multi-objective continuous robot control. *arXiv preprint arXiv:2406.18924*, 2024.

Junjie Song, Yiwen Liu, Dapeng Li, Yin Sun, Shukun Fu, Siqi Chen, and Yuji Cao. Balancing rewards in text summarization: Multi-objective reinforcement learning via hypervolume optimization. *arXiv preprint arXiv:2510.19325*, 2025.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Kevin Tan, Wei Fan, and Yuting Wei. Actor-critics can achieve optimal sample efficiency. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1laMy7jPux.

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. Hypervolume-based multi-objective reinforcement learning. In *International Conference on Evolutionary Multi-Criterion Optimization*, pp. 352–366. Springer, 2013.

Di Wang and Mengqi Hu. Deep deterministic policy gradient with compatible critic network. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4332–4344, 2021.

Yudan Wang, Peiyao Xiao, Hao Ban, Kaiyi Ji, and Shaofeng Zou. Theoretical study of conflict-avoidant multi-objective reinforcement learning. *arXiv preprint arXiv:2405.16077*, 2024.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, et al. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*, 2023.

Mingjing Xu, Peizhong Ju, Jia Liu, and Haibo Yang. Psmgd: Periodic stochastic multi-gradient descent for fast multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21770–21778, 2025.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

Mao Ye and Qiang Liu. Pareto navigation gradient descent: a first-order algorithm for optimization in pareto set. In *Uncertainty in artificial intelligence*, pp. 2246–2255. PMLR, 2022.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond relu to diverse activation functions. *Journal of Machine Learning Research*, 25(35):1–39, 2024.

Xiaoyuan Zhang, Xi Lin, Bo Xue, Yifan Chen, and Qingfu Zhang. Hypervolume maximization: A geometric view of pareto set learning. *Advances in Neural Information Processing Systems*, 36: 38902–38929, 2023.

Zhiyao Zhang, Zhuqing Liu, Xin Zhang, Wen-Yen Chen, Jiyan Yang, and Jia Liu. Multi-objective bilevel learning. *arXiv preprint arXiv:2511.07824*, 2025a.

Zhiyao Zhang, Myeung Suk Oh, FNU Hairi, Ziyue Luo, Alvaro Velasquez, and Jia Liu. Finite-time global optimality convergence in deep neural actor-critic methods for decentralized multi-agent reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=ZcvGJH4ps7.

Zeyu Zhao, Yueling Che, Kaichen Liu, Jian Li, and Junmei Yao. Multi-policy pareto front tracking based online and offline multi-objective reinforcement learning. *arXiv preprint arXiv:2508.02217*, 2025.

Dan Zhou, Jiqing Du, and Sachiyo Arai. Neuroevolutionary diversity policy search for multi-objective reinforcement learning. *Information Sciences*, 657:119932, 2024a.

Tianchen Zhou, FNU Hairi, Haibo Yang, Jia Liu, Tian Tong, Fan Yang, Michinari Momma, and Yan Gao. Finite-time convergence and sample complexity of actor-critic multi-objective reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 61913–61933, 2024b.

Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2810–2818, 2019a.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019b.

APPENDIX

A ADDITIONAL RELATED WORK ON MORL

The strengths of our approach compared to existing methods are highlighted in Table 2.

Table 2: Comparison of Different Algorithms.

Algorithm	Model	Convergence	Rate	Exploration
Qiu et al. (2024)	Tabular	Global	$\mathcal{O}(T^{-\frac{1}{2}})$	✓
Zhou et al. (2024b)	Linear	Stationary	$\mathcal{O}(T^{-1})$	✗
Wang et al. (2024)	Linear	Stationary	$\mathcal{O}(T^{-1})$	✗
Hairi et al. (2025)	Linear	Stationary	$\mathcal{O}(T^{-1})$	✓
Yang et al. (2019)	DNN	NA	-	✗
Chen et al. (2021)‡	DNN	NA	-	✓
This Work	**DNN**	**Global**	$\mathcal{O}(T^{-1})$	✓

Convergence: whether the algorithm converges to Pareto stationary or optimal policies, or if no such guarantee exists. *Rate*: the convergence rate of the algorithm. *Exploration*: whether the algorithm can explore the entire Pareto front. ‡: They study a different setup and do consider multiple preferences.

In addition to the related works discussed in Section 2, we present additional literature in this section. We primarily focus on two more branches of MORL methods and also include works that provide practical guidance for MORL.

Geometric and Pareto analysis-based Approaches. In addition to scalarization-based MORL algorithms, there are works that focus on the geometric properties of the Pareto front and develop approaches inspired by the concept of Pareto optimality (Ye & Liu, 2022; Zhang et al., 2023; Shu et al., 2024; Liu et al., 2025; Zhao et al., 2025; Song et al., 2025). However, these works either do not provide theoretical guarantees for finite-time convergence or only allow convergence to stationary policies.

Evolutionary-based Approaches. Evolutionary approaches (Moriarty et al., 1999; Bai et al., 2023) are also widely used in the field of reinforcement learning. In recent years, several revolutionary MORL algorithms have been proposed (Khadka et al., 2019; Zhou et al., 2024a; Adde et al., 2025; Zhao et al., 2025; Janmohamed & Cully, 2025). These methods adopt a population-based approach to explore the Pareto front in parallel, which improves the efficiency of the algorithms to some extent. However, they remain heuristic and lack foundational theories to precisely characterize their performance.

Measurements in MORL. Several works also provide guidance on the metrics and benchmarks used in MORL (Van Moffaert et al., 2013; Roijers et al., 2018; Hayes et al., 2021; Felten et al., 2023). Metrics such as HyperVolume and the ϵ-metric are also employed in this paper to numerically verify our algorithm.

B THEORETICAL PROOF OF THEOREM 1

Proof. The proof can be divided into three steps. First, we apply the descent lemma to control the dynamic $G_\mu(\theta_t \mid p)$ for each t, leading to an iteration result. Second, we demonstrate that the approximations in Algorithm 1 carefully control each term in the obtained result. Finally, we combine all the components to complete the analysis.

Step A. Apply Descent Lemma and Iterate on $G_\mu(\theta \mid p)$.

We begin by stating the following important lemma.

Lemma 4. *For any $\kappa \geq 0$, and for any $p \in \Delta_m^{++}$, μ, denoting $C(p,\mu) = \frac{M_g}{\sigma} L_J + \mu \log m + \max_i \left(p_i(J_i^{ub} - J_i(\theta^*(p,\mu)) + \frac{\sqrt{\epsilon_{bias}}}{1-\gamma})\right)$, we have:*

$$G_\mu(\theta \mid p) \leq C(p,\mu),$$

where we simply denote $\|\cdot\|_2$ as $\|\cdot\|$ in the sequel.

Proof. According to Kakade & Langford (2002); Ding et al. (2022); Gaur et al. (2024); Zhang et al. (2025b), under Assumptions 4 and 5, for any policy θ and any $i \in [m]$, it holds that:

$$J_i(\theta^*(p,\mu)) - J_i(\theta) \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \frac{M_g}{\sigma}\|\nabla_\theta J_i(\theta)\|.$$

Thus, denoting $h_i(\theta \mid p) = p_i f_i(\theta)$, we can reformulate this result to further obtain:

$$\begin{aligned}
&J_i(\theta^*(p,\mu)) - J_i^{\text{ub}} + J_i^{\text{ub}} - J_i(\theta) \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \frac{M_g}{\sigma}\|\nabla_\theta J_i(\theta)\|, \\
\implies & f_i(\theta) - \frac{M_g}{\sigma}\|\nabla f_i(\theta)\| \leq J_i^{\text{ub}} - J_i(\theta^*(p,\mu)) + \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}, \\
\implies & h_i(\theta \mid p) - \frac{M_g}{\sigma}\|\nabla h_i(\theta \mid p)\| \leq p_i\Big(J_i^{\text{ub}} - J_i(\theta^*(p,\mu)) + \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}\Big).
\end{aligned}$$

On the one hand, according to Assumption 3, we know that $\|\nabla h_i(\theta \mid p)\| \leq p_i L_J, \forall i \in [m], \theta$. On the other hand, the property of "Log-Sum" inequality ensures that $G_\mu(\theta \mid p) \leq \max_i \big(p_i f_i(\theta)\big) + \mu \log m$. Hence, we combine these results to get:

$$\begin{aligned}
G_\mu(\theta \mid p) &\leq \max_i \big(p_i f_i(\theta)\big) + \mu \log m \\
&= \max_i h_i(\theta \mid p) + \mu \log m \\
&\leq \max_i \left(\frac{M_g}{\sigma}\|\nabla h_i(\theta \mid p)\| + p_i\Big(J_i^{\text{ub}} - J_i(\theta^*(p,\mu)) + \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}\Big)\right) + \mu \log m \\
&\leq \frac{M_g}{\sigma} L_J + \mu \log m + \max_i \left(p_i\Big(J_i^{\text{ub}} - J_i(\theta^*(p,\mu)) + \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}\Big)\right) \\
&\leq C(p,\mu),
\end{aligned}$$

holds for any $\kappa \geq 0$, which ends the proof. □

Notably, this is an extension of the *performance difference lemma* (Kakade & Langford, 2002), indicating that the objective function $G_\mu(\theta \mid p)$ derived from smooth WC technique still retains this important property: we can quantify the optimality of any policy θ by evaluating the value $C(p,\mu)$.

Now, we consider the sequence $\{\theta_t\}_{t=0}^{T-1}$, where $\theta_{t+1} = \theta_t - \alpha_t \frac{d_t}{\|d_t\|}$ by Algorithm 1. According to descent lemma, we have:

$$\begin{aligned}
G_\mu(\theta_{t+1} \mid p) &\leq G_\mu(\theta_t \mid p) + \nabla G_\mu(\theta_t \mid p)^\top(\theta_{t+1} - \theta_t) + \frac{L_\mu}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= G_\mu(\theta_t \mid p) - \alpha_t \frac{\nabla G_\mu(\theta_t \mid p)^\top d_t}{\|d_t\|} + \frac{L_\mu \alpha_t^2}{2} \\
&\leq G_\mu(\theta_t \mid p) - \alpha_t\|\nabla G_\mu(\theta_t \mid p)\| + 2\alpha_t\|d_t - \nabla G_\mu(\theta_t \mid p)\| + \frac{L_\mu \alpha_t^2}{2},
\end{aligned}$$

where the last inequality can be derived from the following argument with $v = \nabla G_\mu(\theta_t \mid p) - d_t$ and $d = d_t$:

$$\begin{aligned}
& 0 \leq \|v\| \cdot \|d\| + v^\top d, \\
\iff & \|v\| \cdot \|d\| + \|d\|^2 \leq \|d\|^2 + 2\|v\| \cdot \|d\| + v^\top d, \\
\implies & \|d\| \cdot \|v + d\| \leq \|d\|^2 + 2\|v\| \cdot \|d\| + v^\top d, \\
\iff & -\frac{(v+d)^\top d}{\|d\|} \leq 2\|v\| - \|v + d\|.
\end{aligned}$$

Then, according to Assumption 6, we have the following result:

$$\nabla G_\mu(\theta \mid p) = \sum_{i=1}^{m} \frac{\exp(\frac{h_i(\theta|p)}{\mu})}{\sum_{i'=1}^{m} \exp(\frac{h_{i'}(\theta|p)}{\mu})} \nabla h_i(\theta \mid p) = \Big(H(\theta \mid p)\Big)\lambda(\theta, p, \mu),$$

$$\Longrightarrow \|\nabla G_\mu(\theta \mid p)\| \geq \sigma_{\min}\Big(H(\theta \mid p)\Big)\|\lambda(\theta, p, \mu)\| \geq \frac{\delta_0}{\sqrt{m}},$$

where we denote:

$$\lambda_i(\theta, p, \mu) = \frac{\exp(\frac{h_i(\theta|p)}{\mu})}{\sum_{i'=1}^{m} \exp(\frac{h_{i'}(\theta|p)}{\mu})}, \quad \lambda(\theta, p, \mu) = (\lambda_1(\theta, p, \mu), \ldots, \lambda_m(\theta, p, \mu))^\top.$$

By Lemma 4, we have $G_\mu(\theta \mid p) \leq C(p, \mu)$ for every θ. Therefore, we can get:

$$\left[G_\mu(\theta \mid p) - G_\mu(\theta^*(p, \mu) \mid p) - \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}\right]_+ \leq C(p, \mu) - G_\mu(\theta^*(p, \mu) \mid p).$$

Combining this bound with Assumption 6 yields:

$$\|\nabla G_\mu(\theta \mid p)\| \geq \sqrt{\frac{2\rho}{C(p, \mu) - G_\mu(\theta^*(p, \mu) \mid p)}} \left[G_\mu(\theta \mid p) - G_\mu(\theta^*(p, \mu) \mid p) - \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}\right]_+.$$

If $C(p, \mu) = G_\mu(\theta^*(p, \mu) \mid p)$, then Lemma 4 and the optimality of $\theta^*(p, \mu)$ imply that $G_\mu(\theta \mid p) = G_\mu(\theta^*(p, \mu) \mid p)$ for every θ, and the desired result holds trivially. Otherwise, define $\kappa := \sqrt{\frac{C(p,\mu) - G_\mu(\theta^*(p,\mu)|p)}{2\rho}}$, we obtain:

$$G_\mu(\theta \mid p) - G_\mu(\theta^*(p, \mu) \mid p) \leq \kappa \|\nabla G_\mu(\theta \mid p)\| + \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}.$$

Substituting this inequality into the descent result gives:

$$\begin{aligned} G_\mu(\theta_{t+1} \mid p) - G_\mu(\theta^*(p, \mu) \mid p) \leq \left(1 - \frac{\alpha_t}{\kappa}\right) [G_\mu(\theta_t \mid p) - G_\mu(\theta^*(p, \mu) \mid p)] \\ + 2\alpha_t \|d_t - \nabla G_\mu(\theta_t \mid p)\| + \frac{L_\mu \alpha_t^2}{2} + \frac{\alpha_t}{\kappa} \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma}. \end{aligned} \tag{4}$$

Then, with $\alpha_t = \alpha/t$ and $\alpha > \kappa$, we only need to control the term $\|d_t - \nabla G_\mu(\theta_t \mid p)\|$.

Step B. Control $\|d_t - \nabla G_\mu(\theta_t \mid p)\|$.

For each t, we first add and subtract one term as follows:

$$\|d_t - \nabla G_\mu(\theta_t \mid p)\|^2 \leq \underbrace{2\|d_t - H(\theta_t \mid p)\widehat{\lambda}(\theta_t, p, \mu)\|^2}_{A_t} + \underbrace{2\|H(\theta_t \mid p)\widehat{\lambda}(\theta_t, p, \mu) - \nabla G_\mu(\theta_t \mid p)\|^2}_{B_t}.$$

We also introduce the following notations:

$$\widehat{\lambda}_i(\theta, p, \mu) = \frac{\exp(\frac{p_i(\widehat{J}_i^{\text{ub}} - \widehat{J}_i(\theta))}{\mu})}{\sum_{i'=1}^{m} \exp(\frac{p_{i'}(\widehat{J}_{i'}^{\text{ub}} - \widehat{J}_{i'}(\theta))}{\mu})}, \quad \widehat{\lambda}(\theta, p, \mu) = (\widehat{\lambda}_1(\theta, p, \mu), \ldots, \widehat{\lambda}_m(\theta, p, \mu))^\top.$$

Therefore, for A_t, we have:

$$\begin{aligned} A_t &\overset{\flat}{=} 2\|\sum_{i=1}^{m} \widehat{\lambda}_i(\theta_t, p, \mu) \cdot p_i\big(\nabla J_i(\theta_t) - \widehat{\nabla} J_i(\theta_t)\big)\|^2 \\ &\overset{\dagger}{\leq} 2\sum_{i=1}^{m} \widehat{\lambda}_i(\theta_t, p, \mu)\|p_i\big(\nabla J_i(\theta_t) - \widehat{\nabla} J_i(\theta_t)\big)\|^2 \\ &\overset{\ddagger}{\leq} 2\max_i \|\widehat{\nabla} J_i(\theta_t) - \nabla J_i(\theta_t)\|^2, \end{aligned}$$

where $\flat$ is due to the definition of d_t according to Equation (3), $\dagger$ is due to $\widehat{\lambda}(\theta_t, p, \mu) \in \Delta_m^{++}$ and convexity of $\|\cdot\|^2$, and $\ddagger$ is due to $p \in \Delta_m^{++}$.

In order to further bound A_t, we leverage the following fact according to Cai et al. (2019); Zhang et al. (2025b) to show the optimality of critic after update. Notably, the original results are primarily established for the Q function, whereas it is not difficult to verify that the same arguments also yield parallel results for the V function.

Fact 1 (Cai et al. (2019))**.** *Let locally linear approximated V-function for any parameter W be:*

$$\widehat{V}_0(x; W) = \widehat{V}(x; W(0)) + \langle \nabla_W \widehat{V}(x; W(0)), (W - W(0)) \rangle,$$

and corresponding TD-error be:

$$\delta_0(x, r, x'; W) = \widehat{V}_0(x; W) - r - \gamma \widehat{V}_0(x'; W).$$

If W^* satisfies:

$$\mathbb{E}_{s\sim\nu}\left((\delta_0(x, r, x'; W^*) \cdot \langle \nabla_W \widehat{V}_0(x; W^*)), (W - W^*) \rangle \right) \geq 0, \forall W \in \mathcal{B}(B),$$

where, with a slight abuse of notation, ν here denotes the stationary distribution for state space (under the transition of the restart kernel and some policy π_θ), then we say W^* is a stationary point (since there is no descent direction at W^*). Then, for any A, $W(0)$ and c, there exists some stationary point W^*, and $\widehat{V}_0(\cdot; W^*)$ is the unique, global optimum of the minimization problem for policy θ:

$$\min_W \mathbb{E}_{x\sim\nu(\theta)} \left[(\hat{V}(x; W) - \Pi_{\mathcal{F}_{B,w}} \mathcal{T}^{\pi_\theta} \hat{V}(x; W))^2 \right],$$

where $\mathcal{F}_{B,w} = \{\widehat{V}(x; W(0)) + \langle \nabla_W \widehat{V}(x; W(0)), (W - W(0)) \rangle : W \in \mathcal{B}(B)\}$, and $\Pi_{\mathcal{F}_{B,w}}$ denotes the projection operation to the function class $\mathcal{F}_{B,w}$.

Thus, for each $i \in [m]$, we can reformulate the desired term as:

$$\begin{aligned}
&\|\widehat{\nabla} J_i(\theta_t) - \nabla J_i(\theta_t)\|^2 \\
\leq &\underbrace{3\|\widehat{\nabla} J_i(\theta_t) - \widehat{\nabla} J_i(\theta_t; \delta(W^*_{i,t}))\|^2}_{A_{t,1}} + \underbrace{3\|\widehat{\nabla} J_i(\theta_t; \delta(W^*_{i,t})) - \widehat{\mathrm{Adv}}(\theta_t; \delta(W^*_{i,t}))\|^2}_{A_{t,2}} \\
&+ \underbrace{3\|\widehat{\mathrm{Adv}}(\theta_t; \delta(W^*_{i,t})) - \nabla J_i(\theta_t)\|^2}_{A_{t,3}},
\end{aligned}$$

where:

$$\begin{aligned}
\widehat{\nabla} J_i(\theta_t; \delta(W^*_{i,t})) &:= \frac{1}{M} \sum_{l=0}^{M-1} \delta_{i,t_l}(W^*_{i,t}) \psi_{t_l}, \\
\widehat{\mathrm{Adv}}(\theta_t; \delta(W^*_{i,t})) &:= \mathbb{E}\Big(\mathrm{Adv}(s, a; W^*_{i,t}) \psi_{\theta_t}(s, a) \Big),
\end{aligned}$$

where $\delta_{i,t_l}(W^*_{i,t}) = \widehat{V}(s_{t_l}; W^*_{i,t}) - r_{i,t_{l+1}} - \gamma \widehat{V}(s_{t_{l+1}}; W^*_{i,t})$ adopts the stationary $W^*_{i,t}$ for policy θ_t, and $\psi_{\theta_t}(s, a) = \nabla_\theta \log(\pi_{\theta_t}(s, a))$. Hence, we next show that $A_{t,1}$, $A_{t,2}$, $A_{t,3}$ and B_t can be controlled, respectively.

Step B.1. For $A_{t,1}$, we have:

$$\begin{aligned}
A_{t,1} &= 3\|\frac{1}{M} \sum_{l=0}^{M-1} \delta_{i,t_l} \psi_{t_l} - \frac{1}{M} \sum_{l=0}^{M-1} \delta_{i,t_l}(W^*_{i,t}) \psi_{t_l}\|^2 \\
&= 3\|\frac{1}{M} \sum_{l=0}^{M-1} \Big(\delta_{i,t_l} - \delta_{i,t_l}(W^*_{i,t}) \Big) \psi_{t_l}\|^2 \\
&\leq 3 \max_l \|\big(\delta_{i,t_l} - \delta_{i,t_l}(W^*_{i,t}) \big) \psi_{t_l}\|^2 \\
&\leq 3M_g^2 \max_l \big(\delta_{i,t_l} - \delta_{i,t_l}(W^*_{i,t}) \big)^2,
\end{aligned}$$

where the last inequality is due to Assumption 4. For each $i \in [m]$, we can further get:

$$\begin{aligned}
&\left|\delta_{i,t_l} - \delta_{i,t_l}(W^*_{i,t})\right| \\
=&\left|\widehat{V}(s_{t_l}; W_{i,t}) - \gamma\widehat{V}(s_{t_{l+1}}; W_{i,t}) - \widehat{V}(s_{t_l}; W^*_{i,t}) + \gamma\widehat{V}(s_{t_{l+1}}; W^*_{i,t})\right| \\
\leq&\left|\widehat{V}(s_{t_l}; W_{i,t}) - \widehat{V}(s_{t_l}; W^*_{i,t})\right| + \gamma\left|\widehat{V}(s_{t_{l+1}}; W_{i,t}) - \widehat{V}(s_{t_{l+1}}; W^*_{i,t})\right|,
\end{aligned}$$

where the equality is due to the definition of TD-errors, and the inequality comes from the triangle inequality. Then, we take expectation, and follow the parallel results in Cai et al. (2019); Zhang et al. (2025b) to get that:

$$\mathbb{E}\left[\left|\delta_{i,t_l} - \delta_{i,t_l}(W^*_{i,t})\right|\right] = \mathcal{O}\left((BD^{\frac{5}{2}}K^{-\frac{1}{4}} + B^{\frac{4}{3}}w^{-\frac{1}{12}}D^4)\cdot\log^{\frac{3}{2}} w\log^{\frac{1}{2}} K\right),$$

holds with probability at least $1-\exp(-\Omega(\log^2 w))$, when selecting $\beta = 1/\sqrt{K}$, $w = \Omega(d^3D^{-\frac{11}{2}})$, $B = \Theta(w^{\frac{1}{32}}D^{-6})$, and $K = \Omega(D^4)$. Then, this implies:

$$\mathbb{E}(A_{t,1}) = \mathcal{O}\left((B^2D^5K^{-\frac{1}{2}} + B^{\frac{8}{3}}w^{-\frac{1}{6}}D^8)\cdot\log^3 w\log K\right). \tag{5}$$

Step B.2. For $A_{t,2}$, we have:

$$\begin{aligned}
\frac{A_{t,2}}{3} &= \Big\|\frac{1}{M}\sum_{l=0}^{M-1}\delta_{i,t_l}(W^*_{i,t})\psi_{t_l} - \mathbb{E}\Big(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\Big)\Big\|^2 \\
&= \frac{1}{M^2}\sum_{l=0}^{M-1}\Big\|\delta_{i,t_l}(W^*_{i,t})\psi_{t_l} - \mathbb{E}\Big(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\Big)\Big\|^2 \\
&+ \frac{1}{M^2}\sum_{u\neq v}\langle\delta_{i,t_u}(W^*_{i,t})\psi_{t_u} - \mathbb{E}\left(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\right), \delta_{i,t_v}(W^*_{i,t})\psi_{t_v} - \mathbb{E}\left(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\right)\rangle.
\end{aligned}$$

We take expectation on both sides. Then, We first control the first term in the last equation as follows. For each $l \in \{0,\ldots,M-1\}$, we have:

$$\begin{aligned}
&\mathbb{E}\Big\|\delta_{i,t_l}(W^*_{i,t})\psi_{t_l} - \mathbb{E}\Big(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\Big)\Big\|^2 \\
\leq&2\mathbb{E}\Big[\|\delta_{i,t_l}(W^*_{i,t})\psi_{t_l}\|^2 + \|\mathbb{E}\left(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\right)\|^2\Big] \\
\leq&2\mathbb{E}\Big[|\delta_{i,t_l}(W^*_{i,t})|^2\cdot\|\psi_{t_l}\|^2\Big] + 2\mathbb{E}\Big[|\text{Adv}(s,a;W^*_{i,t})|^2\cdot\|\psi_{\theta_t}(s,a)\|^2\Big] \\
\leq&2M_g^2\mathbb{E}|\delta_{i,t_l}(W^*_{i,t})|^2 + 2M_g^2\mathbb{E}|\text{Adv}(s,a;W^*_{i,t})|^2 \\
\leq&4M_g^2\max_l\mathbb{E}|\delta_{i,t_l}(W^*_{i,t})|^2 \\
=&4M_g^2\Big((\frac{1+\gamma}{1-\gamma}+1)r_{\max} + 2\widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\right)\Big)^2 \\
=&\mathcal{O}(1),
\end{aligned}$$

where the second last equation is due to the definition of TD-errors and Lemma 2, and the last equation holds because of the bounded reward. Then, we consider the second term. Without loss of generality, we assume $u < v$. Besides, we consider taking expectations conditioned on the filtration $\mathcal{F}_t$, where $\mathcal{F}_t$ denotes the samples up to iteration t. According to Hairi et al. (2022); Zhou et al. (2024b) the following results hold due to Assumption 2:

$$\begin{aligned}
&\mathbb{E}\Big[\langle\delta_{i,t_u}(W^*_{i,t})\psi_{t_u} - \mathbb{E}\left(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\right), \delta_{i,t_v}(W^*_{i,t})\psi_{t_v} - \mathbb{E}\left(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\right)\rangle\Big|\mathcal{F}_t\Big] \\
=&2\Big((\frac{1+\gamma}{1-\gamma}+1)r_{\max} + 2\widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\right)\Big) \\
&\cdot\mathbb{E}\left[\Big\|\mathbb{E}\left(\text{Adv}(s_{t,l_v},a_{t,l_v};W^*_{i,t})\psi_{\theta_t}(s,a)\big|\mathcal{F}_{t,l_u}\right) - \mathbb{E}\left(\text{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\right)\Big\|\Big|\mathcal{F}_t\right] \\
=&2M_g\Big((\frac{1+\gamma}{1-\gamma}+1)r_{\max} + 2\widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\right)\Big)^2\cdot\eta\rho^{v-u}.
\end{aligned}$$

Hence, we can further get:

$$\mathbb{E}(A_{t,2}) = \mathcal{O}\left(\frac{1}{M}\right) + \mathcal{O}\left(\frac{1}{M^2}\cdot M\frac{\eta\rho}{1-\rho}\right) = \mathcal{O}\left(\frac{1}{M}\right). \tag{6}$$

Step B.3. For $A_{t,3}$, according to Lemma 2, we have:

$$\begin{aligned}
A_{t,3} &= \Big\|\mathbb{E}\Big(\mathrm{Adv}(s,a;W^*_{i,t})\psi_{\theta_t}(s,a)\Big) - \mathbb{E}\Big(\mathrm{Adv}_{\theta_t,i}(s,a)\psi_{\theta_t}(s,a)\Big)\Big\|^2 \\
&= \Big\|\mathbb{E}\Big[\big(\mathrm{Adv}(s,a;W^*_{i,t}) - \mathrm{Adv}_{\theta_t,i}(s,a)\big)\psi_{\theta_t}(s,a)\Big]\Big\|^2 \\
&\le M_g^2\Big(\mathbb{E}\big|\mathrm{Adv}(s,a;W^*_{i,t}) - \mathrm{Adv}_{\theta_t,i}(s,a)\big|\Big)^2 \\
&= M_g^2\Big(\mathbb{E}\big|\gamma\mathbb{E}(\widehat{V}(s;W^*_{i,t})) - \widehat{V}(s;W^*_{i,t}) - \gamma\mathbb{E}(V_{\theta_t,i}(s)) + V_{\theta_t,i}(s)\big|\Big)^2 \\
&\le 4M_g^2\Big(\mathbb{E}\big|\widehat{V}(s;W^*_{i,t}) - V_{\theta_t,i}(s)\big|\Big)^2 \\
&\le 4M_g^2\mathbb{E}\big(\widehat{V}(s;W^*_{i,t}) - V_{\theta_t,i}(s)\big)^2 \\
&= \widetilde{\mathcal{O}}\left(w^{-\frac{4}{d}}D^{-\frac{4}{d}}\right).
\end{aligned} \tag{7}$$

Step B.4. As for B_t, we can obtain:

$$\begin{aligned}
B_t &= 2\|H(\theta_t \mid p)\big(\widehat{\lambda}(\theta_t,p,\mu) - \lambda(\theta_t,p,\mu)\big)\|^2 \\
&= 2\|\sum_{i=1}^m \big(\widehat{\lambda}_i(\theta_t,p,\mu) - \lambda_i(\theta_t,p,\mu)\big)\cdot p_i\nabla J_i(\theta_t)\|^2 \\
&\overset{\dagger}{\le} 2\sum_{i=1}^m p_i\|\big(\widehat{\lambda}_i(\theta_t,p,\mu) - \lambda_i(\theta_t,p,\mu)\big)\cdot\nabla J_i(\theta_t)\|^2 \\
&\overset{\ddagger}{\le} 2L_J^2\sum_{i=1}^m p_i\big(\widehat{\lambda}_i(\theta_t,p,\mu) - \lambda_i(\theta_t,p,\mu)\big)^2 \\
&\le 2L_J^2\max_i\big(\widehat{\lambda}_i(\theta_t,p,\mu) - \lambda_i(\theta_t,p,\mu)\big)^2.
\end{aligned}$$

where † is due to $p\in\Delta_m^{++}$ and convexity of $\|\cdot\|^2$, and ‡ is due to Assumption 3.

This implies that, for any $i\in[m]$, we need to consider $\big(\widehat{\lambda}_i(\theta_t,p,\mu) - \lambda_i(\theta_t,p,\mu)\big)^2$. To this end, we first consider the following bias arising from approximations:

$$\begin{aligned}
&\mathbb{E}\big|\widehat{J}_i(\theta_t) - J_i(\theta_t)\big| \\
=&\mathbb{E}\left|\widehat{V}(s_0;W_{i,t}) - V_{\theta_t,i}(s_0)\right| \\
\le&\mathbb{E}\left|\widehat{V}(s_0;W_{i,t}) - \widehat{V}(s_0;W^*_{i,t})\right| + \mathbb{E}\left|\widehat{V}(s_0;W^*_{i,t}) - V_{\theta_t,i}(s_0)\right| \\
=&\mathcal{O}\left((BD^{\frac{5}{2}}K^{-\frac{1}{4}} + B^{\frac{4}{3}}w^{-\frac{1}{12}}D^4)\cdot\log^{\frac{3}{2}}w\log^{\frac{1}{2}}K\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\right).
\end{aligned}$$

Let $\mathcal{J}_i(\theta) = p_i\frac{J_i^{\mathrm{ub}} - J_i(\theta)}{\mu}$ and $\widehat{\mathcal{J}}_i(\theta) = p_i\frac{J_i^{\mathrm{ub}} - \widehat{J}_i(\theta)}{\mu}$. Then, we can get:

$$\mathbb{E}\big|\widehat{\mathcal{J}}_i(\theta_t) - \mathcal{J}_i(\theta_t)\big| = \mathcal{O}\left((BD^{\frac{5}{2}}K^{-\frac{1}{4}} + B^{\frac{4}{3}}w^{-\frac{1}{12}}D^4)\cdot\log^{\frac{3}{2}}w\log^{\frac{1}{2}}K\mu^{-1}\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\mu^{-1}\right).$$

Then, we denote $\mathcal{J}(\theta) = (\mathcal{J}_1(\theta),\ldots,\mathcal{J}_m(\theta))^\top$, and $\widehat{\mathcal{J}}(\theta) = (\widehat{\mathcal{J}}_1(\theta),\ldots,\widehat{\mathcal{J}}_m(\theta))^\top$. For convenience, we also denote $\phi_i(\mathcal{J}(\theta)) = \frac{\exp(\mathcal{J}_i(\theta))}{\sum_j\exp(\mathcal{J}_j(\theta))}$. Then, according to Mean value theorem, we know

that there exists some $c \in \mathbb{R}^m$, such that:

$$\begin{aligned}\left|\phi_i(\mathcal{J}(\theta)) - \phi_i(\widehat{\mathcal{J}}(\theta))\right|^2 &= \left|\nabla\phi_i(c)^\top\left(\mathcal{J}(\theta) - \widehat{\mathcal{J}}(\theta)\right)\right|^2 \\ &\leq \|\nabla\phi_i(c)\|_1^2 \|\mathcal{J}(\theta) - \widehat{\mathcal{J}}(\theta)\|_\infty^2 \\ &\leq \frac{1}{4}\max_i \left|\widehat{\mathcal{J}}_i(\theta_t) - \mathcal{J}_i(\theta_t)\right|^2,\end{aligned}$$

which implies that:

$$\begin{aligned}&\mathbb{E}\Big(\widehat{\lambda}_i(\theta_t, p, \mu) - \lambda_i(\theta_t, p, \mu)\Big)^2 \\ =&\mathbb{E}\Big(\phi_i(\mathcal{J}(\theta_t)) - \phi_i(\widehat{\mathcal{J}}(\theta_t))\Big)^2 \\ \leq&\mathbb{E}\Big[\frac{1}{4}\max_i\Big(\widehat{\mathcal{J}}_i(\theta_t) - \mathcal{J}_i(\theta_t)\Big)^2\Big] \\ \leq&\frac{1}{4}\mathbb{E}\Big[\sum_{i=1}^m\Big(\widehat{\mathcal{J}}_i(\theta_t) - \mathcal{J}_i(\theta_t)\Big)^2\Big] \\ =&\mathcal{O}\left((B^2D^5K^{-\frac{1}{2}} + B^{\frac{8}{3}}w^{-\frac{1}{6}}D^8)\cdot\log^3 w\log K\mu^{-2}m\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{4}{d}}D^{-\frac{4}{d}}\mu^{-2}m\right),\end{aligned}$$

holds with probability at least $1 - \exp(-\Omega(\log^2 w))$ when selecting parameters as shown before. These results indicate that:

$$\mathbb{E}(B_t) = \mathcal{O}\left((B^2D^5K^{-\frac{1}{2}} + B^{\frac{8}{3}}w^{-\frac{1}{6}}D^8)\cdot\log^3 w\log K\mu^{-2}m\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{4}{d}}D^{-\frac{4}{d}}\mu^{-2}m\right). \quad (8)$$

Step C. Complete the Proof.

Combining Equations (5) to (8), we know that the following result holds with probability at least $1 - \exp(-\Omega(\log^2 w))$ when selecting parameters as shown before:

$$\begin{aligned}\|d_t - \nabla G_\mu(\theta_t \mid p)\| = &\mathcal{O}\left((BD^{\frac{5}{2}}K^{-\frac{1}{4}} + B^{\frac{4}{3}}w^{-\frac{1}{12}}D^4)\cdot\log^{\frac{3}{2}} w\log^{\frac{1}{2}} K\mu^{-1}m^{\frac{1}{2}}\right) \\ &+ \widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\mu^{-1}m^{\frac{1}{2}}\right) + \mathcal{O}\left(M^{-\frac{1}{2}}\right).\end{aligned}$$

Then, we substitute this back to Equation (4) to obtain the following result:

$$\begin{aligned}&G_\mu(\theta_{t+1} \mid p) - G_\mu(\theta^*(p,\mu) \mid p) \\ \leq&\frac{1}{t}G_\mu(\theta_2 \mid p) + \frac{2\alpha}{t}\sum_{\tau=2}^t\|d_\tau - \nabla G_\mu(\theta_\tau \mid p)\| + \frac{L_\mu\alpha^2}{2t} + \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} \\ =&\mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}\left(\sqrt{\epsilon_{\text{bias}}}\right) + \mathcal{O}\left(M^{-\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\mu^{-1}m^{\frac{1}{2}}\right) \\ &+ \mathcal{O}\left((BD^{\frac{5}{2}}K^{-\frac{1}{4}} + B^{\frac{4}{3}}w^{-\frac{1}{12}}D^4)\cdot\log^{\frac{3}{2}} w\log^{\frac{1}{2}} K\mu^{-1}m^{\frac{1}{2}}\right)\end{aligned}$$

Therefore, be selecting $\alpha_t = \frac{\alpha}{t}$, $\alpha \geq \max\{1, \frac{M_g L_J\sqrt{m}}{\sigma\delta_0} + \frac{\mu\log m\sqrt{m}}{\delta_0}\}$, $\beta = 1/\sqrt{K}$, $w = \Omega(d^3D^{-\frac{11}{2}})$, $B = \Theta(w^{\frac{1}{32}}D^{-6})$, and $K = \Omega(D^4)$, with probability at least $1 - \exp^{-\Omega(\log^2 w)}$, Algorithm 1 has the following global convergence guarantee for any $p \in \Delta_m^{++}$:

$$\begin{aligned}&\mathbb{E}\Big[G_\mu(\theta_T \mid p) - G_\mu(\theta^*(p,\mu) \mid p)\Big] \\ =&\mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\sqrt{\epsilon_{\text{bias}}}\right) + \mathcal{O}\left(M^{-\frac{1}{2}}\right) + \mathcal{O}\left(w^{-\frac{2}{d}}D^{-\frac{2}{d}}\mu^{-1}m^{\frac{1}{2}}\right) \\ &+ \widetilde{\mathcal{O}}\left(w^{\frac{1}{32}}D^{-\frac{7}{2}}K^{-\frac{1}{4}}\mu^{-1}m^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(w^{-\frac{1}{24}}D^{-4}\mu^{-1}m^{\frac{1}{2}}\right),\end{aligned}$$

which ends the proof. □

Based on Theorem 1, which controls the converted smoothed-WC problem, we can also conclude the following corollary for the original MORL problem.

Corollary 1. *Let Δ_T denote the complete upper bound in Theorem 1 and $p_{\min} = \min_i p_i > 0$. Define the Pareto-domination gap as $D(\theta) = \sup_{\theta'}[\min_i J_i(\theta') - J_i(\theta)]$. Then, θ_T is $\Delta_T/p_{\min}$-weak Pareto optimal for the original MORL problem.*

Proof. If a policy θ' improves every objective of θ by at least ϵ, then the coordinatewise monotonicity of the log-sum-exp operator gives:

$$G_\mu(\theta'|p) \le G_\mu(\theta|p) - p_{\min}\epsilon.$$

It follows that:

$$\begin{aligned}
D(\theta) &= \sup_{\theta'} \left[\min_i J_i(\theta') - J_i(\theta)\right], \\
\text{(for any } \delta > 0\text{, exists } \theta_\delta) &\implies \min_i [J_i(\theta_\delta) - J_i(\theta)] \ge D(\theta) - \delta, \\
\text{(for } 0 < \delta < D(\theta)) &\implies G_\mu(\theta_\delta|p) \le G_\mu(\theta|p) - p_{\min}(D(\theta) - \delta), \\
\text{(def. of minimizer)} &\implies G_\mu(\theta^*(p,\mu)|p) \le G_\mu(\theta_\delta|p) \le G_\mu(\theta|p) - p_{\min}(D(\theta) - \delta), \\
\text{(let } \delta \text{ go to 0)} &\implies G_\mu(\theta^*(p,\mu)|p) \le G_\mu(\theta|p) - p_{\min} D(\theta).
\end{aligned}$$

Arranging terms, dividing both sides by $p_{\min}$, and using Theorem 1, we have:

$$\mathbb{E}[D(\theta_T)] \le \frac{\mathbb{E}[G_\mu(\theta|p) - G_\mu(\theta^*(p,\mu)|p)]}{p_{\min}} \underbrace{\le}_{\text{Thm. 1}} \frac{\Delta_T}{p_{\min}}.$$

Thus, Theorem 1 ensures that θ_T is $\Delta_T/p_{\min}$-approximately weakly Pareto optimal for the original MORL problem, and the convergence rate remains the same. □

C Setups and Additional Results of Numerical Experiments

In this section, we provide the detailed setups of our numerical experiments along with some additional results.

C.1 Simulation on Recommendation Systems

1) Detailed Setup. We conduct experiments on the Kuairand dataset (Gao et al., 2022), an unbiased sequential recommendation dataset collected from the recommendation logs of a video-sharing mobile app. It provides rich feedback of 12 distinct signals (e.g. click, like, view time, follows, comments) together with timestamps, user and item features, and over 30 side-features. Its unbiased exposure mechanism makes it especially useful for evaluating debiasing and causal recommendation methods.

To compare with existing algorithms, we evaluate algorithm performances on optimizing 7 main feedback signals: click, like, follow, comment, forward, dislike, view time, and compare our approach with several baselines (Chen et al., 2021; Zhou et al., 2024b; Hairi et al., 2022). To ensure a fair comparison, for DNN based methods (i.e., Chen et al. (2021), our work), we implement them using the same network architecture of a 3-layer perceptron each with ReLU as activation function for both critic and actor networks. For methods utilizing linear approximation, i.e., Zhou et al. (2024b); Hairi et al. (2025), we keep both critic and actor networks as single linear layer for linear approximation. In addition, for methods with preference vector as input (i.e., Hairi et al. (2025), our work), we set a unified preference vector for all objectives. We note here that the reward optimization among objectives could still be biased even with unified preference vector, given that different feedback signals have very different density in this data, e.g., the density of signal "forward" is 0.076% in all 7 signals so there is very limited customer feedback can be learned. Finally, we benchmark all methods on metric normalized capped importance sampling (NCIS) (Zou et al., 2019a).

Notably, the use of ReLU and Sigmoid activations can be interpreted as follows. (1) ReLU networks possess *positive scaling invariance*, which may introduce degenerate parameter directions and make Assumption 5 difficult to satisfy. Sigmoid networks avoid this particular issue and are therefore more

compatible with our theoretical assumptions. We thus use Sigmoid in our analysis. (2) On the other hand, we adopt ReLU in the experiments because it generally provides better empirical optimization performance. More importantly, ReLU-based MLPs with depth D and width w can approximate any Lipschitz continuous function with a vanishing error rate of $\widetilde{\mathcal{O}}(w^{-2/d}D^{-2/d})$. Sigmoid achieves the *same* order of universal approximation error as ReLU-based MLPs. Specifically, the set of MLPs with width no greater than $3w$ and depth no greater than $2D$, i.e., $\mathcal{NN}(3w, 2D; \mathsf{Sigmoid})$ is dense in $\mathcal{NN}(w, D; \mathsf{ReLU})$. This implies that, for the Lipschitz continuous function, **the universal approximation error of a** Sigmoid**-based** $(3w, 2D)$**-MLPs also vanishes at the same rate of** $\widetilde{\mathcal{O}}(w^{-2/d}D^{-2/d})$ (Shen et al., 2022; Zhang et al., 2024).

The preference vector used for the preference-based algorithms (Ours, MOCHA) shown in Table 1 is $p = \frac{\lambda}{||\lambda||}$ where $\lambda = [10.0, 1.0, 1.0, 1.0, 0.01, 0.1, 10.0]^\top$. We select this "seemingly random" preference vector to achieve a relatively balanced performances across all objectives, since, as mentioned above, the distribution of the dataset is highly biased. This is also the preference vector used to obtain the reference point in Figure 2. Moreover, the (unnormalized) preference vectors used in Figure 2 are listed in Table 3.

Table 3: Preference vectors used in Figure 2.

P_{click}	$[100.0, 1.0, 1.0, 1.0, 0.01, 0.1, 10.0]^\top$
P_{like}	$[10.0, 100.0, 1.0, 1.0, 0.01, 0.1, 10.0]^\top$
P_{follow}	$[10.0, 1.0, 100.0, 1.0, 0.01, 0.1, 10.0]^\top$
P_{comment}	$[10.0, 1.0, 1.0, 100.0, 0.01, 0.1, 10.0]^\top$
P_{forward}	$[10.0, 1.0, 1.0, 1.0, 100.0, 0.1, 10.0]^\top$
P_{dislike}	$[10.0, 1.0, 1.0, 1.0, 0.01, 100.0, 10.0]^\top$
$P_{\text{watchtime}}$	$[10.0, 1.0, 1.0, 1.0, 0.01, 0.1, 100.0]^\top$

2) Additional Results. We also provide the numerical results on the metrics of hypervolume and ϵ-metric Hayes et al. (2021) as follows:

Table 4: hypervolume results in recommendation system (comparison with baselines).

Algorithm	Behavior-Clone	MOAC	MOCHA	PDPG	**Ours**
hypervolume (↑)	0.054	0.094	0.083	0.138	**0.223**

The performance on hypervolume for our algorithm and the baselines is shown in Table 4. It is worth noting that some baselines are preference-free methods, indicating that they can only converge to a **single** Pareto solution, rather than multiple solutions across the Pareto front. Therefore, to ensure a fair comparison with these methods, the hypervolume degenerates to "boxvolume" here. Specifically, we compute the area "between" the *single* solution and a fixed reference point for every algorithm: (i) for preference-free algorithms, the single solution corresponds to the one the algorithm converges to, and (ii) for preference-based algorithms, the solution is the one listed in Table 1, i.e., based on the preference $p = \frac{\lambda}{||\lambda||}$ where $\lambda = [10.0, 1.0, 1.0, 1.0, 0.01, 0.1, 10.0]^\top$. Clearly, our algorithm outperforms all other baselines (and allowing varying preference vectors leads to the union of such "boxvolumes", which will be larger than 0.223), which directly confirms the effectiveness of our approach.

We also provide the numerical results on the ϵ-metric for the preference vectors listed in Table 3. As shown in Figure 3, except for the diagonal elements, all values are greater than 1, indicating that the trade-offs among conflicting objectives are well balanced by selecting distinct preference vectors. This further validates the systematic Pareto exploration capability of our algorithm.

C.2 Experiments on Bi-Objective Robotic Simulation

1) Detailed Setup. In MuJoCo-Walker-2d-v5 environment (Felten et al., 2023), we aim to control a walking robot (walker) to move forward. The basic setups are detailed as follows:

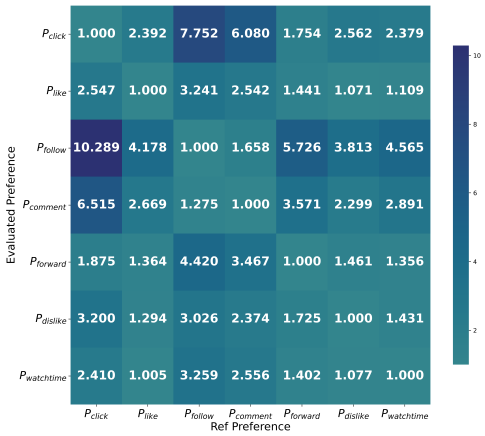
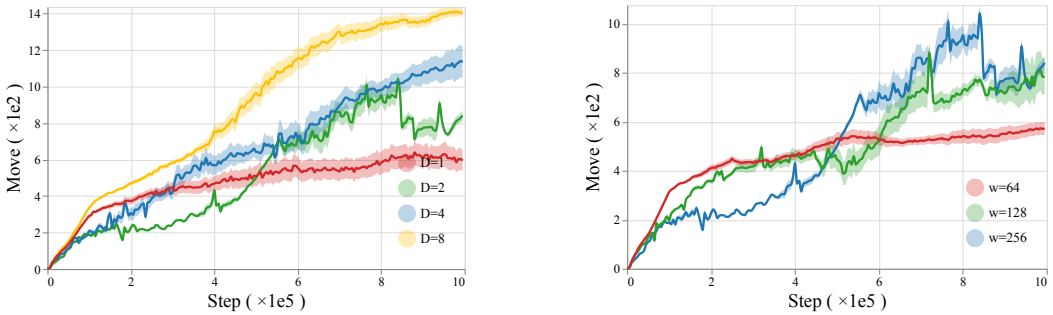


Figure 3: ϵ-metric performances for different preference vectors.

- **Episode.** In each episode, there are at most $T = 500$ time steps. Every episode ends once the walker falls down (referred to as the walker becoming "unhealthy") or when the maximum number of steps is reached, and a new episode is then initialized. Every experiment consists of a total of $1,000,000$ time steps.
- **State Space.** The state space has a dimension of 17, with each component representing either the position or velocity of a walker's body part. Among these, the first dimension, the position of the "z-coordinate", determines the walker's health: it's healthy only if this value lies within the interval $[0.8, 1.0]$.
- **Action Space.** The action space has a dimension of 6, with each component representing torque added to a specific body part of the walker. The transition kernel follows the laws of mechanics.
- **Reward Signals.** We define the following quantities: 1) "Health" outputs 1 if the robot remains standing, and 0 otherwise; 2) "Forward" denotes the velocity along the forward axis; and 3) "Cost" is proportional to $\|a\|_2^2$, i.e., the squared norm of the taken action a. We then consider two types of reward signals as follows. First, ***Move=Health+Forward*** represents the velocity along the forward axis, encouraging the walker to move forward quickly without falling. Second, ***Control=Health-Cost*** implies that the controller is also supposed to minimize the interference. Obviously, these two kinds of reward signals conflict with each other, motivating us to formulate this robotic simulation task as a bi-objective problem, in which we aim to maximize both objective functions.
- **Preference vectors.** In the experiments, we first randomly sample various preference vectors across the standard simplex Δ_2^+ (Xu et al., 2020; Felten et al., 2024). After obtaining the Pareto front based on these preferences, we observe that the distribution of the resulting points is highly uneven across the front. Therefore, for the sparser regions, where the outputs for the related preferences are more sensitive, we manually select additional preferences to more precisely characterize the shape of the Pareto front. For instance, if the points obtained by $p_1 = [0.9, 0.1]^\top$ and $p_1 = [0.8, 0.2]^\top$ are far apart, we randomly sample additional preference vectors between them, such as $p' = [0.85, 0.15]^\top$, to fill the gap between them. For the ablation studies shown in Figure 4, we fix the preference vector to $p = [0.9, 0.1]^\top$, and modify other hyperparameters.
- **Others.** We also enumerate the default setting of other aforementioned parameters here. The discounted factor is set to $\gamma = 0.99$. The DNNs we utilized are with width of 256 and depth of 2. The smoothing parameter is set to $\mu = 0.05$. The upper bound is set to $\mathbf{J}^{\text{ub}} = \{2000, 1000\}$. The learning rates are set to $\alpha_t = \alpha = 3 \times 10^{-4}$, and $\beta = 10^{-3}$.

In addition to the Pareto exploration results, we also investigate how the parameters impact the performance of our approach. The detailed investigation setups are listed below:

- Depth D. We consider the DNN with different depths: $D \in \{1, 2, 4, 8\}$.
- Width w. We apply the DNN with different widths: $w \in \{64, 128, 256\}$.



(a) Varying depth D.

(b) Varying width w.

Figure 4: Performance under different depth D and hidden width H under preference [0.9,0.1]. Each curve is smoothed using a moving average over 500 steps.

2) Additional Results. We repeat each experiments with 3 different random seeds, and plot the mean and 95% confidence interval. Figure 4 provides the ablation studies by showing the forward reward curves under the default setup while varying a single hyperparameter.

As illustrated in Figure 4a, the reward returns consistently increase as the depth D of the DNN increases. This suggests that the deeper DNN achieves better performance, which aligns with our theoretical results. Figure 4b considers the effect of the width w of the DNN. While larger widths yield slightly better performance, the differences across settings are relatively not significant, which is also consistent with our theoretical analysis.

In addition, we also numerically verify the validity of some supporting assumptions. Specifically, in this robotic simulation experiment, we store the upper bound of the score function's ℓ_2-norm. Correspondingly, the upper bound of the score function's ℓ_2-norm is: $\|\nabla_\theta \log \pi_\theta(a|s)\| \leq 1.51 \times 10^2$.