Multi-Objective Reinforcement Learning with Max-Min Criterion: A Game-Theoretic Approach

Woohyeon Byeon[1] **Giseung Park**[2] **Jongseong Chae**[1] **Amir Leshem**[3]

Youngchul Sung[1]*

Abstract

In this paper, we propose a provably convergent and practical framework for multi-objective reinforcement learning with max-min criterion. From a game-theoretic perspective, we reformulate max-min multi-objective reinforcement learning as a two-player zero-sum regularized continuous game and introduce an efficient algorithm based on mirror descent. Our approach simplifies the policy update while ensuring global last-iterate convergence. We provide a comprehensive theoretical analysis on our algorithm, including iteration complexity under both exact and approximate policy evaluations, as well as sample complexity bounds. To further enhance performance, we modify the proposed algorithm with adaptive regularization. Our experiments demonstrate the convergence behavior of the proposed algorithm in tabular settings, and our implementation for deep reinforcement learning significantly outperforms previous baselines in many MORL environments.

1 Introduction

Reinforcement Learning (RL) focuses on sequential decision-making in Markov Decision Processes (MDPs), which have been extensively studied both theoretically and practically. However, in several practical decision-making problems such as autonomous vehicles, resource allocation and communication, multiple objectives should be simultaneously optimized instead of a conventional single objective [27, 23, 24, 60]. These scenarios motivate the development of Multi-Objective Reinforcement Learning (MORL), an extension of RL in which the reward function yields a vector rather than a scalar. In recent years, MORL has seen significant progresses [61, 50, 31, 37, 20]. A widely-adopted approach to MORL is the utility-based approach [46, 22], which is formulated as follows: for a utility function u (also called scalarization function), solve $\max_\pi u(V_1^\pi, \dots, V_K^\pi)$, where π is a policy and V_k^π is the value function of the k-th objective. As long as the utility function is non-decreasing in the sense of Pareto dominance, the optimal policy for $u(V_1^\pi, \dots, V_K^\pi)$ is also Pareto optimal [46].

Although many previous works on MORL considered the weighted sum utility function [61, 8, 31], the weighted sum utility function is not the desired design metric in many cases, especially when considering the application of MORL to resource allocation where fairness is an important issue. For example, consider scheduling of cloud computing resources, in which a computing job is typically divided into multiple subtasks and each subtask is performed by a different resource and completed subtasks are combined to finish the overall job [47, 57]. In this case, the overall time for finishing the

[1] School of Electrical Engineering, KAIST, Republic of Korea, [2] University of Toronto Robotics Institute, Canada, [3] Faculty of Engineering, Bar-Ilan University, Israel. *Correspondence to: Youngchul Sung <yc-sung@kaist.ac.kr>.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

job is determined by the maximum of subtask completion times. Consider another example of traffic signal light control at an intersection combining multiple roads. One can consider the waiting time at each road and may want to minimize the maximum of the waiting times of all roads for fairness on the drivers at all roads rather than to minimize the sum of the waiting times. Hence, the min-max (or equivalently max-min with negation of objectives) criterion naturally arises in many real-world resource allocation problems as well as other application domains [13, 42, 63, 34, 29, 21, 17], where fairness is important. (Please see Appendix B for comparison with other scalarizartion criteria including proportional fairness.)

In this paper, we consider this **max-min MORL**, where the utility function is the minimum function: $u(x_1, \dots, x_K) = \min_{k=1,\dots,K} x_k$ and the resulting problem is given by $\max_\pi \min_{k=1,\dots,K} V_k^\pi$. With the increasing relevance of max-min MORL, Park et al. [37] recently proposed a **model-free** algorithm for max-min MORL through a convex formulation to circumvent the non-differentiability of min operation. Although their work provides a model-free algorithm to solve exact max-min MORL in contrast to previous approximation methods [19, 39, 50] or model-based approaches [14] to max-min MORL, their algorithm suffers from high memory requirement and computational costs and only guarantees average-iterate convergence. In this paper, we further advance max-min MORL and propose a fast algorithm for max-min MORL with last-iterate convergence and significantly improved efficiency in computing time and required memory by reformulating max-min MORL as a *two-player zero-sum game*.

Our contributions are summarized as follows:

• We propose a single-loop algorithm for entropy-regularized max-min MORL by leveraging appropriate regularization to derive a closed-form update.

• Our algorithm can be viewed as a primal-dual algorithm and we provide a theoretical analysis of the proposed algorithm in the tabular case, including global last-iterate convergence, iteration complexity, and sample complexity.

• We demonstrate convergence behavior through numerical simulations in the tabular case and show that our deep reinforcement learning implementation significantly outperforms existing baselines in several MORL environments including a realistic traffic signal control task.

2 Background

Multi-objective Markov decision process A multi-objective Markov decision process (MOMDP) extends a standard MDP by incorporating a vector-valued reward function. Formally, a MOMDP is defined as $\langle S, A, P, \mu, \mathbf{r}, \gamma \rangle$ with state space S, action space A, transition $P : S \times A \to \Delta(S)$, where $\Delta(X)$ denotes the set of probability distributions on set X, initial state distribution $\mu \in \Delta(S)$ and discount factor $\gamma \in (0, 1)$. Unlike single-objective MDPs, the reward function $\mathbf{r} : S \times A \to \mathbb{R}^K$ is vector-valued with $\mathbf{r}(s, a) = (r_1(s, a), \dots, r_K(s, a))$, where K is the number of objectives. An agent sequentially interacts with an environment by taking an action from its stationary policy $\pi : S \to \Delta(A)$. Unlike single-objective settings, the agent seeks to optimize all components of the reward vector simultaneously in some manner.

The multi-objective state value of a policy π is defined as $\mathbf{V}^\pi = \mathbb{E}_{\mu,\pi} \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t)\right] \in \mathbb{R}^K$, and V_k^π denotes the k-th element of $\mathbf{V}^\pi$. The multi-objective state value with entropy regularization is defined as $\mathbf{V}_\tau^\pi = \mathbb{E}_{\mu,\pi} \left[\sum_t \gamma^t \left(\mathbf{r}(s_t, a_t) - \tau \log \pi(a_t|s_t) \mathbf{1}_K\right)\right] \in \mathbb{R}^K$, where τ is a regularization coefficient, $\mathbf{1}_K$ is a K-dimensional vector all of which entries are 1, and $V_{k,\tau}^\pi$ denotes the k-th element of vector $\mathbf{V}_\tau^\pi$. Throughout this paper, $w = (w(1), \dots, w(K)) \in \Delta^K$ denotes a weight vector, where Δ^K is the $(K-1)$-simplex, i.e., $\Delta^K = \{w | w(i) \geq 0 \, \forall i, \; w(1) + \cdots + w(K) = 1\}$. For a weight vector w, we define the weighted value as $V_w^\pi = \langle w, \mathbf{V}^\pi \rangle$ and the weighted soft value as $V_{w,\tau}^\pi = \langle w, \mathbf{V}_\tau^\pi \rangle$, where $\langle \cdot, \cdot \rangle$ denotes inner product. The soft values can be decomposed as $V_{k,\tau}^\pi = V_k^\pi + \tau \tilde{H}(\pi), \, \forall k$ and $V_{w,\tau}^\pi = V_w^\pi + \tau \tilde{H}(\pi)$, where $\tilde{H}(\pi) = \mathbb{E}_{\mu,\pi} \left[-\sum_t \gamma^t \log \pi(a_t|s_t)\right]$ is the expected cumulative entropy of policy π. A summary of notations is provided in Appendix A.

Max-min MORL Since the value function for MORL is vector-valued, optimizing the value function $\mathbf{V}^\pi$ over π cannot be defined as in single-objective RL. Thus, **max-min MORL** considers the following max-min optimization: $\max_\pi \min_{k=1,\dots,K} V_k^\pi$. However, it is known that max-min MORL may yield the indeterminacy of its solution and entropy regularization resolves the possible indeterminacy of max-min MORL solution [37]. So, in this paper, instead of max-min MORL, we can consider the following **entropy-regularized max-min MORL** problem:

$$\max_{\pi} \min_{k=1,\dots,K} [\, \underbrace{V_k^{\pi} + \tau \tilde{H}(\pi)}_{=V_{k,\tau}^{\pi}} \,] \tag{1}$$

which converges to max-mix MORL as $\tau \to 0$. It can be shown that the value gap between the entropy unregularized and regularized problems is proportional to the coefficent τ. Due to the non-linearity of the $\min$ operator, max-min MORL or entropy-regularized max-min MORL cannot be solved directly using standard RL methods, requiring a specialized framework such as [37] or approximate approaches [19, 39, 50].

Mirror descent and natural policy gradient Mirror descent (MD) [36] is a general optimization framework that subsumes popular algorithms such as gradient descent and the multiplicative weights update [6]. It has been widely applied in learning dynamics across various games [1, 65, 7, 59]. Formally, the update rule for MD can be expressed in primal domain as

$$x_{t+1} = \max_{x \in \mathcal{X}} \lambda \langle \nabla f(x_t), x \rangle - D_{\psi}(x, x_t), \tag{2}$$

where $f(x)$ is the objective function, λ is a step size, and D_{ψ} is the Bregman divergence induced by a continuously differentiable and strictly convex function ψ, defined as $D_{\psi}(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$, $\forall x, y \in \mathcal{X}$. MD basically implements steepest descent in a manifold of which Riemannian metric is given by the Hessian of D_{ψ}, capturing relevant geometry of the optimization variable space. In reinforcement learning, the natural policy gradient (NPG) can be interpreted as a MD variant for MDPs with the Fisher information metric [26, 36, 48].

3 The Proposed Method

Our goal in this paper is to solve the general problem (1). We here develop a single-loop algorithm by interpreting entropy-regularized max-min MORL (1) as learning a Nash equilibrium in a *two-player zero-sum game* and exploiting theory of learning in games.

3.1 Max-min versus min-max

In this section, we derive Theorem 3.1, which serves as the initial step for our reformulation of the entropy-regularized max-min MORL problem (1) as a two-player zero-sum game.

Our key idea begins with the fact that the entropy-regularized max-min MORL problem (1) is equivalent to the following problem [37]:

$$\text{Minimization Reformulation} : \min_{w \in \Delta^K} V_{w,\tau}^{*} \tag{3}$$

where $V_{w,\tau}^{*}$ is the optimal soft value under the scalarized reward $\langle w, \mathbf{r} \rangle$. Here, the equivalence of problems means that both problems have the same optimal value and an optimal solution of one problem yields an optimal solution of the other. The equivalence was first shown in [37] and a brief explanation of this equivalence is provided in Appendix C for readers' convenience.

The next enabling step is that the problem (1) involving minimization over discrete variable k can be converted to minimization over continuous variable $w \in \Delta^K$ as follows [9]:

$$\max_{\pi} \min_{k=1,\dots,K} V_{k,\tau}^{\pi} = \max_{\pi} \min_{w \in \Delta^K} \langle w, \mathbf{V}_{\tau}^{\pi} \rangle, \tag{4}$$

where $\langle w, \mathbf{V}_{\tau}^{\pi} \rangle = \sum_{k=1}^{K} w(k) V_{k,\tau}^{\pi}$, and $w(k)$ is the k-th element of weight vector w. This can easily be seen by considering the case in which we have a single minimum $V_{k',\tau}^{\pi}$ for some k'. In this case, the minimizing w is given by $(0, \cdots, 0, 1, 0, \cdots, 0)$ with 1 at the k'th position. Note that while the two problems in (4) yield the same optimal value, the minimizers are not always in one-to-one correspondence. When the minimum is attained at multiple k indices, the corresponding optimal weights w form a convex combination of those indices and lie on the boundary of the simplex. Theorem 3.1 below formally bridges this gap by proving equivalence of optimal solutions.

Furthermore, due to $V_{w,\tau}^{*} \triangleq \max_{\pi} \langle w, \mathbf{V}_{\tau}^{\pi} \rangle$ by its definition, the problem (3) can be rewritten as

$$\min_{w \in \Delta^K} \max_{\pi} \langle w, \mathbf{V}_{\tau}^{\pi} \rangle. \tag{5}$$

Combining these relations, we have the following equivalence between the max-min problem (4) (=(1)) and the min-max problem (5) (=(3)), i.e.,

$$\max_{\pi} \min_{w\in\Delta^K} \langle w, \mathbf{V}^{\pi}_{\tau}\rangle = \min_{w\in\Delta^K} \max_{\pi} \langle w, \mathbf{V}^{\pi}_{\tau}\rangle. \tag{6}$$

Note that the max-min value is not equal to the min-max value in general. Since value functions in RL are non-concave in policy even with direct parameterization [2], the equality (6) cannot be directly induced from the minimax theorem [45].

From this special relation that the max-min value equals the min-max value in our case, the existence of a saddle point is guaranteed in Proposition F.1 in Appendix F.1. Once the existence of a saddle point is guaranteed, Theorem 3.1 establishes the equivalence of this saddle point (i.e., Nash equilibrium) in (6) and the solution of entropy-regularized max-min MORL (1); i.e., it suffices to find a Nash equilibrium in (6) to solve entropy-regularized max-min MORL.

Theorem 3.1. *Let $(\bar{\pi}, \bar{w})$ be a saddle point in (6). Then, $\bar{\pi}$ is a solution to the entropy-regularized max-min MORL* (1).

Proof: See Appendix F.2.

Now, based on Theorem 3.1, we will reformulate entropy-regularized max-min MORL as learning a Nash equilibrium achieving equality of (6) in the two-player zero-sum game whose payoff functions are given by $\langle w, \mathbf{V}^{\pi}_{\tau}\rangle$ and $-\langle w, \mathbf{V}^{\pi}_{\tau}\rangle$ for efficient algorithm construction.

3.2 Two-plaer zero-sum regularized continuous game formulation

We first define a two-player zero-sum continuous game $\mathcal{G}$ as follows. There are two players; $Learner$ and $Adversary$. The player $Learner$ corresponds to the RL agent in the MOMDP who learns the policy parameter θ. We define the strategy space of $Learner$ as $X_{Learner} = \Theta \subset \mathbb{R}^d$, which is the space of policy parameters in the MOMDP. In other words, $Learner$ takes continuous strategy $\theta \in \Theta$, which indicates the parameter of policy. The player $Adversary$ has the strategy space $X_{adv} = \Delta^K$ and takes continuous strategy $w \in \Delta^K$. The utility functions of the two players are defined as

$$u^{\mathcal{G}}_{Learner}(\theta, w) = \langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle = -u^{\mathcal{G}}_{Adv}(\theta, w). \tag{7}$$

In this game, $Learner$ tries to maximize its utility $\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle$ and $Adversary$ tries to maximize its utility $-\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle$, i.e., minimize $\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle$. In a NE of this game, max-min and min-max values become the same and (6) is achieved to yield our policy solution.

Instead of learning a NE in $\mathcal{G}$, however, we solve the following regularized game $\mathcal{RG}$, which is judiciously designed to our advantage for efficient algorithm construction. The regularized game $\mathcal{RG}$ has the same players and strategy spaces as the original game $\mathcal{G}$. Instead, $\mathcal{RG}$ has utility functions regularized from the utility functions of $\mathcal{G}$. Formally, the utility functions of $\mathcal{RG}$ are defined as

$$u^{\mathcal{RG}}_{Learner}(\theta, w) = \underbrace{\langle w, \mathbf{V}^{\pi_\theta}\rangle + \tau\tilde{H}(\pi_\theta)}_{=\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle} - \tau_w H(w) = -u^{\mathcal{RG}}_{Adv}(\theta, w), \tag{8}$$

where $\tilde{H}(\pi) = \mathbb{E}_{\mu,\pi}\left[-\sum_t \gamma^t \log \pi(a_t|s_t)\right]$ as already defined, $H(w) := -\sum_{k=1}^{K} w(k)\log w(k)$, and τ and τ_w are positive coefficients.

Now, let us explain why we solve the new regularized game $\mathcal{RG}$ instead of solving $\mathcal{G}$. For $Learner$, the main reason for adding entropy regularization $\tilde{H}(\pi_\theta)$ is to avoid the indeterminacy problem as already explained [37]. In short, stationary deterministic policies are insufficient for max-min MORL since there exists an MOMDP that has only a stochastic optimal policy for max-min MORL [46].

For $Adversary$, adding $-H(w)$ regularization, $Adversary$ minimizes $\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle - \tau_w H(w)$. That is, $Adversary$ minimizes $\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle$ while increasing the entropy $H(w)$. Note that $Learner$ maximizes $\min_{w\in\Delta^K}\langle w, \mathbf{V}^{\pi_\theta}_{\tau}\rangle$ with its policy π_θ. Without $-H(w)$ regularization, w achieving $\min_{w\in\Delta^K}\langle w, \mathbf{V}^{\pi}_{\tau}\rangle$ would be one-hot vector focusing only on the worst dimension of $\mathbf{V}^{\pi_\theta}_{\tau}$. However, with $-H(w)$ regularization, the elements of the w vector will be spread out to incorporate multiple objective dimensions for $Learner$'s optimization. This speeds up learning and we can achieve last-iterate convergence rather than average-iterate convergence, which will be proven in Section 4. In addition, unlike other strongly convex regularizations such as squared l_2-norm, our choice of negative entropy regularization enables a *closed-form solution* for the MD update of w shown in (12) in Section 3.3. This closed-form update due to our $-H(w)$ regularization significantly simplifies the optimization process and reduces computational overhead.

3.3 The ERAM algorithm: Entropy-regularized adversary for max-min MORL

We propose **ERAM** (Entropy-Regularized Adversary for Max-min MORL), an efficient algorithm to solve max-min MORL via a game-theoretic perspective. Based on the reformulation of the problem as a two-player zero-sum game, we leverage equilibrium learning methods and introduce a variant of the MD algorithm to find a Nash equilibrium in $\mathcal{RG}$.

First, we update the policy parameter with the MD objective on $u^{\mathcal{RG}}_{Learner}$ with step size η as follows:

$$\theta_{t+1} = \arg\max_{\theta} \{\eta\langle\nabla_\theta \, u^{\mathcal{RG}}_{Learner}(\theta_t, w_t), \, \theta\rangle - D_\psi(\theta, \theta_t)\}. \tag{9}$$

Note that the objective in (9) can be viewed as the Lagrangian of the optimization of the linear approximation of $Learner$'s utility around current θ_t (i.e., $u^{\mathcal{RG}}_{Learner}(\theta_t, w_t) + \langle\nabla_\theta \, u^{\mathcal{RG}}_{Learner}(\theta_t, w_t), \, \theta - \theta_t\rangle$) under a constraint on $D_\psi(\theta, \theta_t)$. Here, we choose the convex function ψ for the Bregman divergence to be the negative Shannon entropy to make the Bregman divergence the KL-divergence, which yields the Fisher information matrix (FIM) as Hessian matrix and is relevant to statistical manifolds: That is, π_θ is a probablity distribution with coordinate θ, and $\{\pi_\theta\}$ forms a statistical manifold where FIM is an invariant metric [5]. Then, the optimization with such conservatism in RL leads to natural policy gradient (NPG) [5, 36, 26, 48]. The NPG update rule is given by [11]

$$\theta_{t+1} = \theta_t + \eta F^\dagger(\theta_t) \sum_{k=1}^{K} w_t(k) \nabla_\theta V^{\pi_{\theta_t}}_{k,\tau}, \tag{10}$$

where $F(\theta) = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}}[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^T]$ is the FIM, η is a step size for NPG, and † denotes the Moore–Penrose pseudo-inverse [25]. In the discrete tabular case, with softmax policy parameterization, i.e., $\pi_\theta(a|s) = \frac{e^{\theta_{sa}}}{\sum_{a'} e^{\theta_{sa'}}}$, $\theta \in \mathbb{R}^{|S||A|}$, which is general enough to cover non-negative stochastic policies, we have the closed-form for NPG update [11]:

$$\pi_{\theta_{t+1}}(a|s) = \frac{1}{Z_\pi(t,s)} (\pi_{\theta_t}(a|s))^\alpha \exp\left(\frac{1-\alpha}{\tau} Q^{\pi_{\theta_t}}_{w_t,\tau}(s,a)\right), \tag{11}$$

where $\alpha = 1 - \frac{\eta\tau}{1-\gamma}$, and $Z_\pi(t,s)$ is a normalization constant. In the case of linear function approximation for π_θ, NPG can easily be implemented with compatible function approximation [55]. In the case of general nonlinear neural network parameterization, the computation of FIM is difficult but the NPG update for θ can readily be replaced with TRPO [48] or PPO [49] of which purpose is to solve such divegence-constrained policy optimization. We will use PPO for deep implementation.

For the update of the strategy w of $Adversary$, we use a variant of MD objective on $u^{\mathcal{RG}}_{Adv}$ with step size λ as follows:

$$w_{t+1} = \arg\max_{w\in\Delta^K} \lambda\langle\nabla_w(-\langle w, \mathbf{V}^{\pi_{\theta_t}}\rangle)|_{w=w_t}, \, w\rangle + \lambda\tau_w H(w) - D_\psi(w, w_t). \tag{12}$$

Note that $\tau\tilde{H}(\pi_\theta)$ is irrelevant to w update, and $\tau_w H(w)$ is outside of the gradient. Again, we choose the Bregman divergence to be the KL divergence, considering $\{w\}$ forms a weight simplex Δ^K, a manifold identical to the probability simplex of K outcomes. With this choice, we have the following closed-form solution to (12) [9]:

$$w_{t+1} = \mathrm{softmax}\left(-\frac{1-\beta}{\tau_w}\mathbf{V}^{\pi_{\theta_t}} + \beta \log w_t\right) \tag{13}$$

where $\beta = \frac{1}{\lambda\tau_w+1}$. In this way, we circumvent any complicated iterative procedure to update w_t. Note that we could have derived another closed-form solution to w_{t+1} for conventional MD on $u^{\mathcal{RG}}_{Adv}$, which yields the same solution as in (13) with $\beta = 1 - \lambda\tau_w$. We prefer the modified MD update because the resulting closed-form solution employs $\beta = \frac{1}{\lambda\tau_w+1}$, which is guaranteed to lie in $(0,1)$ for any choice of hyperparameters λ and τ_w. We note that (11) and (13) together lead to an NE of the game $\mathcal{RG}$, as shown in Section 4.

3.4 The ARAM algorithm: Adaptively-regularized adversary for max-min MORL

Although ERAM provides an efficient model-free algorithm based on PPO and closed-form update of w_t, the performance can further be improved by detailing the w_t update. We here develop Adaptively

Regularized Adversary for Max-min MORL (**ARAM**), a variant of ERAM. In ARAM, we adopt an adaptive regularization method that extends the standard entropy regularization. Note that the regularizer $H(w)$ in the $Adversary$'s update rule (12) can be expressed as the KL divergence from the uniform reference weight $\frac{1}{K}\mathbf{1}_K$, i.e., $H(w) = -D_{\text{KL}}\left(w \,\middle\|\, \frac{1}{K}\mathbf{1}_K\right) + \log K$. Thus, increasing entropy is equivalent to minimizing the distance from the uniform reference. In ARAM, we replaced the uniform reference $\frac{1}{K}\mathbf{1}_K$ with a dynamically computed vector $c \in \Delta^K$, leading to the regularizer $-D_{\text{KL}}(w\|c)$ instead of $H(w)$, where c captures the correlation between each reward component and the worst-performing objective in the previous iteration. That is, the i-th element of the reference vector c is obtained as

$$c_i = \text{softmax}(\mathbb{E}_{s,a}[r_i(s,a)r_{i'}(s,a)]),\ i = 1, \cdots, K, \tag{14}$$

where i' is the index of the worst-performing objective at the previous iteration batch and the expectation is replaced with sample expectation. Thus, ERAM minimizes $\langle w, \mathbf{V}_\tau^{\pi_\theta}\rangle$ while trying to keep w close to $\frac{1}{K}\mathbf{1}_K$, i.e., considering all objective dimensions equally. On the other hand, ARAM minimizes $\langle w, \mathbf{V}_\tau^{\pi_\theta}\rangle$ while putting more emphasis on poorly-performing objective dimensions but not solely on the worst dimension, enabling joint optimization of multiple objective dimensions unlike previous GGF-PPO [50] which optimizes only the worst dimension at each batch. Again, a closed-form formula for the w update in ARAM can be derived and more on ARAM is provided in Appendix D.

Summarizing the above, the pseudo-codes of the proposed algorithms are in Appendix E. Our source code is provided at `https://github.com/whbyeon/ERAM-ARAM`.

4 Theoretical Analysis

In this section, we provide theoretical analysis of ERAM with respect to convergence, iteration complexity, and sample complexity under the assumption of tabular MOMDP in which the closed-form update (11) for θ is available with the softmax policy parameterization of $\pi_\theta(a|s)$, while leaving theoretical analysis of more complicated ARAM as a possible future work. Note that the term "global convergence" means "convergence regardless of initial condition", and the term "last-iterate convergence of sequence x_t to x^*" means "$x_t \to x^*$ as $t \to \infty$", whereas the term "average-iterate convergence" means "$\frac{1}{T}\sum_{t=1}^T x_t \to x^*$ as $t \to \infty$", which does not imply $x_t \to x^*$ in general.

Note that for (11) and (13) we need to know the action value function since the state value function can readily be obtained from the action value function by summing over actions. We consider two cases regarding the knowledge of the action value function, exact and approximate policy evaluation cases, for proof of convergence.

Global convergence with exact policy evaluation: Theorem 4.1 shows the last-iterate convergence of Algorithm 1 when policy evaluation is exact.

Theorem 4.1. *Let $\{\theta_t\}_t$ and $\{w_t\}_t$ are the sequences generated by Algorithm 1 and let $\pi_t = \pi_{\theta_t}$. Then, the optimality gaps satisfy the following:*

$$\|\log\pi^* - \log\pi_t\|_\infty \le C_1[\rho(\eta,\lambda)]^t \tag{15}$$

$$\|w^* - w_t\|_\infty \le C_2[\rho(\eta,\lambda)]^t \tag{16}$$

$$\|Q_{w^*,\tau}^{\pi^*} - Q_{w_t,\tau}^{\pi_t}\|_\infty \le C_3[\rho(\eta,\lambda)]^t \tag{17}$$

for some C_1, C_2, C_3, where $0 < \rho(\eta,\lambda) \le 1 - \frac{\epsilon^2}{2} < 1$ with $\eta = \frac{\epsilon(1-\gamma)}{\tau}$, $\tau_w \ge \frac{12K(\max_{s,a,k}|r_k(s,a)|+\tau\log|A|)^2}{\tau(1-\gamma)^4} > 0$ and $\epsilon \in (0,\epsilon_0)$ for some ϵ_0.

Proof: See Appendix G.

For last-iterate convergence of policy and weight, the step size of weight-update $\lambda = O(\epsilon^2)$ should have a smaller scale than the step size of NPG $\eta = O(\epsilon)$. We note that if $\eta = O(\epsilon)$, then it suffices for λ to have a smaller scale at least $O(\epsilon^p),\ p > 1$. Our choice $\lambda = O(\epsilon^2)$ is one possible choice that satisfies this condition. In an intuitive sense, the policy is required to be updated faster than the weight.

Corollary 4.2. *Let the desired accuracy error tolerance be denoted by ϵ_{acc}. To achieve $\|\log\pi^* - \log\pi_t\|_\infty \le \epsilon_{acc}$, $\|w^* - w_t\|_\infty \le \epsilon_{acc}$, $\|Q_{w^*,\tau}^{\pi^*} - Q_{w_t,\tau}^{\pi_t}\|_\infty \le \epsilon_{acc}$ with $\eta = \frac{\epsilon(1-\gamma)}{\tau}$, $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)}$, softmax policy and exact policy evaluation, Algorithm 1 requires at most $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon_{acc}})$ iterations.*

Global convergence with approximate policy evaluation: When exact policy evaluation is not available, our algorithm can be adapted to use approximate value estimates with bounded error. We establish that the last-iterate convergence guarantee still holds under this relaxed setting.

Theorem 4.3. *Assume that the estimated values $\widehat{Q}^{\pi}_{w,\tau}$ and $\widehat{Q}^{\pi}_{k}$ satisfy $\|\widehat{Q}^{\pi}_{w,\tau} - Q^{\pi}_{w,\tau}\|_\infty < \delta$ and $\|\widehat{Q}^{\pi}_{k} - Q^{\pi}_{k}\|_\infty < \delta$ for any π, w, k. Let $\{\theta_t\}_t$ and $\{w_t\}_t$ be the sequences generated by Algorithm 2, a modified version of Algorithm 1 for approximate policy evaluation provided in Appendix E, and let $\pi_t = \pi_{\theta_t}$. Then, the optimality gaps satisfy the following:*

$$\|\log \pi^* - \log \pi_t\|_\infty \le \widehat{C}_1[\widehat{\rho}(\eta, \lambda)]^t + \widehat{D}_1\delta/\epsilon^2 \tag{18}$$

$$\|w^* - w_t\|_\infty \le \widehat{C}_2[\widehat{\rho}(\eta, \lambda)]^t + \widehat{D}_2\delta/\epsilon^2 \tag{19}$$

$$\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty \le \widehat{C}_3[\widehat{\rho}(\eta, \lambda)]^t + \widehat{D}_3\delta/\epsilon^2 \tag{20}$$

where $0 < \widehat{\rho}(\eta, \lambda) < 1 - \frac{\epsilon^2}{2} < 1$ for $\epsilon \in (0, \epsilon_0)$ with the same condition in Theorem 4.1 except for the values of $\widehat{D}_1, \widehat{D}_2, \widehat{D}_3$ and ϵ_0.

Proof is similar to that of Theorem 4.1 and is provided in Appendix H.

Corollary 4.4. *Let the accuracy error tolerance be denoted by ϵ_{acc}. Assume that the estimation error δ is sufficiently small to satisfy $\delta \le \frac{\epsilon^2\epsilon_{acc}}{\widehat{D}_i}$, $i = 1, 2, 3$. To achieve $\|\log \pi^* - \log \pi_t\|_\infty \le 2\epsilon_{acc}$, $\|w^* - w_t\|_\infty \le 2\epsilon_{acc}$, and $\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty \le 2\epsilon_{acc}$ with $\eta = \frac{\epsilon(1-\gamma)}{\tau}$, $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)}$, softmax policy with approximate policy evaluation, Algorithm 2 requires at most $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon_{acc}})$ iterations. Furthermore, with estimation error $\delta = O(\epsilon^2\epsilon_{acc})$ under Assumptions J.1 and J.2 provided in Appendix J, by employing fresh samples for the policy evaluation for each objective at every iteration and taking the union bound for all objectives and all $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon_{acc}})$ iterations, Algorithm 2 requires at most $\tilde{O}(\frac{K}{(1-\gamma)^3\epsilon^4\epsilon_{acc}^2}) \times O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon_{acc}}) = \tilde{O}(\frac{K}{(1-\gamma)^3\epsilon^6\epsilon_{acc}^2})$ samples per each state-action pair.*

Remark 4.5. *It can be shown that the difference between the optimal value function induced the regularized game $\mathcal{RG}$ and that of the unregularized counterpart (i.e., $\mathcal{RG}$ with $\tau = \tau_w = 0$) is upper-bounded linearly in the regularization coefficients τ and τ_w. This implies that the optimal solution of the regularized game yields a value function close to that of the original max–min MORL problem (i.e.,* (1) *with $\tau = 0$), as long as the regularization coefficients are sufficiently small.*

Proofs of Corollaries 4.2 and 4.4 are provided in Appendix I, and details for the sample complexity analysis are provided in Appendix J.

5 Related Works

Max-min MORL The utility-based approach is a key aspect of MORL, as it incorporates fairness and reflects user preferences [46, 22]. To capture broader notions of fairness, recent studies have adopted non-linear utility functions. Cousins et al. [14] considered a model-based approach to fair RL but this method is applicable to small finite problems. Fan et al. [19] analyzed fairness through the lens of the Nash social welfare function, while Peng and Fain [39] proposed a reward-aware value iteration framework for general non-linear welfare functions, including the $\min$ operator. These works focused on optimizing the objective $\mathbb{E}_\pi\left[\min_k \sum_t \gamma^t r_k(s_t, a_t)\right]$, which is not the true min value $\min_k \mathbb{E}_\pi\left[\sum_t \gamma^t r_k(s_t, a_t)\right]$, to simplify the problem. Siddique et al. [50] studied fair policy learning in MORL using the generalized Gini social welfare function, which includes max-min fairness as a special case. GGF-DQN, their DQN-based method, optimizes again the surrogate objective $\mathbb{E}_\pi\left[\min_k \sum_t \gamma^t r_k(s_t, a_t)\right]$ due to the difficulty in constructing a Bellman operator under the non-linearity of the $\min$ operator. GGF-PPO, their policy-based method, performs PPO update every iteration batch based on the current minimum objective value dimension. Note that if we remove $H(w)$ by setting $\tau_w = 0$ and remove $D_\psi(w, w_t)$ in our w update (12), then the strategy w of $Adversary$ is the one-hot vector with element 1 at the minimum dimension of $\mathbf{V}^{\pi_\theta}_\tau$ and this case corresponds to GGF-PPO. Hence, our work can be considered as a generalization of GGF-PPO. However, when w is constrained to switching among one-hot vectors as in GGF-PPO, only average-iterate convergence is guaranteed [6]. Note that our method has $H(w)$ encouraging to consider multiple dimensions simultaneously and $D_\psi(w, w_t)$ preventing rapid jumps in w for last-iterate convergence. Performance improvement will be shown in Section 6.

For max-min fairness based on direct optimization of $\min_k \mathbb{E}_\pi \left[\sum_t \gamma^t r_k(s_t, a_t)\right]$, Park et al. [37] proposed a theoretical framework based on primal-dual convex programming for maximum-entropy reinforcement learning. To implement this framework, they introduced a model-free double-loop algorithm: the inner loop computes a stochastic zeroth-order gradient estimator using Gaussian smoothing, and the outer loop performs projected gradient descent with the estimated gradient. The complexity of this method is far beyond that of our current method.

Game-theoretic learning with regularization Regularization techniques have been widely used to compute Nash equilibria in game-theoretic settings. APMD [1] analyzes MD under smooth monotone game assumptions, which do not hold in our RL-based setting due to the non-monotonicity of the value gradients. In addition, APMD applies MD to all agents, whereas we combine RL-oriented NPG for the learner and MD for the adversary, which better suits deep RL implementations. Perolat et al. [40] studied two-player zero-sum Markov games based on replicator dynamics and provided an asymptotic convergence analysis using Lyapunov techniques. Aggarwal et al. [3] focused on regularized linear-quadratic games and established non-asymptotic convergence guarantees. Zeng et al. [64] proposed a regularized gradient descent–ascent method for two-player zero-sum Markov games and provided non-asymptotic analysis. In contrast to these approaches, our method addresses a heterogeneous setting with one RL learner and a non-RL adversary, and is specifically tailored to max-min criterion for MORL. We also provide non-asymptotic convergence for our method. For a detailed discussion of zero-sum Markov game literature, we refer the reader to Appendix K.

Primal-dual methods and distributionally-robust RL Several constrained RL methods adopt a primal–dual framework to solve constrained MDPs, where a Lagrangian function is formulated and an alternating optimization procedure is applied to maximize over the policy and minimize over the Lagrange multipliers [10, 33, 32, 17, 18]. Although our algorithm can be viewed as one instance of primal-dual algorithms, our work differs from these approaches in that the weight w lies in a probability simplex, whereas Lagrange multipliers of these works reside in the non-negative quadrant. In addition, due to $w \in \Delta^K$, we employ entropy regularization $H(w)$ instead of the $\|w\|^2$ regularization used in Müller et al. [33], which allows us to obtain a closed-form solution for the w-update without requiring projected gradient descent used in Efroni et al. [18], Müller et al. [33].

Distributionally-robust reinforcement learning (DR-RL) typically addresses transition uncertainty [52, 16, 30]. Analogously, in reward-uncertain MDPs [44], one can define an uncertainty set in the reward space and apply a max–min formulation to achieve robustness. Our setting corresponds to a finite uncertainty set $\{r_1, \ldots, r_K\}$, where the regularization term captures such internalized reward uncertainty, in a manner similar to how DR-RL handles transition uncertainty. For infinite uncertainty sets, DR-RL could potentially be extended to analyze robustness in terms of a distribution over the reward space.

6 Experiments

6.1 Numerical results in tabular setting

We empirically demonstrate the convergence of our algorithm in tabular MOMDP settings. We considered three types of tabular MOMDPs with different sizes: $(|S|, |A|, K) = (2, 2, 2)$, $(3, 3, 6)$, and $(4, 4, 4)$. As the suboptimality measure, we used the Nash gap, a common metric in game theory [53, 38], defined as $\text{Nash\_Gap}(\theta, w) = (\max_{\theta'} \langle w, \mathbf{V}^{\pi_{\theta'}} \rangle - \langle w, \mathbf{V}^{\pi_\theta} \rangle) + (\langle w, \mathbf{V}^{\pi_\theta} \rangle - \min_{w'} \langle w', \mathbf{V}^{\pi_\theta} \rangle)$. This measures how far the strategy tuple (θ, w) is from a Nash equilibrium in the unregularized game.

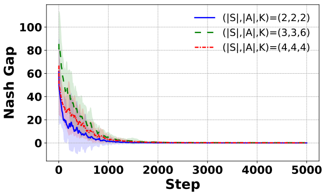


Figure 1: Nash gap

Figure 1 shows that the Nash gap of Algorithm 2 decreases quickly over time in all three tabular settings. Each curve represents the average over 50 randomly generated instances, with shaded areas showing standard deviation. We also present the convergence behavior of ARAM with approximate policy evaluation in the tabular settings in Figure 7 in Appendix L.

6.2 Experimental results in traffic signal control

To evaluate the effectiveness of our algorithm in real-world multi-objective problems, we conducted experiments in the traffic signal control simulation environment [4]. At a four-road intersection, the agent controlled traffic signals based on traffic state information and received a reward vector composed of the negative total waiting times. Rewards were defined either per road (4 objectives) or per lane (16 objectives). We considered three scenarios: Base-4, Asym-4, and Asym-16, which differ in the number of objectives and traffic flow symmetry, as explained in Appendix M.1.

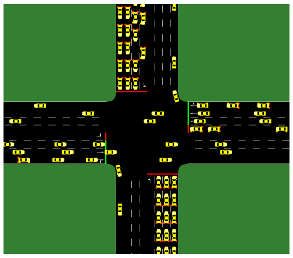


Figure 2: Traffic signal control environment [4]

We compared our method against GGF-DQN, GGF-PPO [50], Park et al. [37], and Avg-DQN. Avg-DQN optimizes the average reward $\frac{1}{K}\sum_k r_k(s,a)$, reflecting simple sum approaches. GGF-DQN optimizes a surrogate objective, $\mathbb{E}_\pi\left[\min_k \sum_t \gamma^t r_k(s_t,a_t)\right]$, which yields a lower bound on the true max-min objective $\min_k \mathbb{E}_\pi[\sum_t \gamma^t r_k(s_t,a_t)]$ via Jensen's inequality. GGF-PPO performs a PPO update on the objective of minimum value at each iteration. In contrast, the method of Park et al. [37] directly solves the target problem (1) without relying on a surrogate, using projected gradient descent with a smoothed gradient estimate via Gaussian smoothing. We evaluated the max-min performance as the minimum of the empirical return vector, i.e., $\min_k \hat{R}_k = \frac{1}{N}\sum_{i=1}^{N}\sum_t \gamma^t r_k^i(s_t,a_t)$, averaged over $N = 32$ episodes and five random seeds. (Simulation details are provided in Appendix M.1.)

Environments	ARAM	ERAM	Park et al. [37]	GGF-PPO	GGF-DQN	Avg-DQN
Base-4	**-1160**	-1387	-1681	-1731	-1838	-2774
Asym-4	**-2696**	-2732	-3510	-3501	-3053	-4245
Asym-16	**-15043**	-17334	-23663	-21663	-17792	-27499

Table 1: Max-min performance in traffic signal control. Bold: best; underline: second-best.

Table 1 reports the max-min performance across the traffic signal control environments. ARAM consistently outperforms all baselines across all environments. ERAM achieves comparable results while maintaining architectural simplicity, and both methods directly optimize the target objective without relying on surrogate losses.

More experimental results on other environments such as the species conservation environment [51], MO-Reacher environment [20], and Four-Room environment [20] are provided in Appendix N, showing the superior max-min performance of our algorithms.

6.3 Complexity comparison

Convergence behavior Our method guarantees last-iterate convergence, unlike Park et al. [37], which ensures only average-iterate convergence. Figure 3 illustrates this difference in simple MOMDPs with two states, two actions, and two objectives. We present results on three representative randomly generated MOMDPs to ensure visual clarity; the same color is used to indicate the same MOMDP instance across the two plots. In each case, ERAM consistently approaches the true optimal value, up to a gap determined by the regularization coefficient, while the baseline exhibits oscillatory behavior.

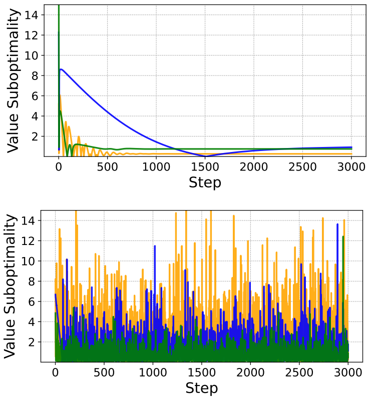


Figure 3: Convergence comparison in random tabular MOMDPs: ERAM (top) vs. Park et al. [37] (bottom). Colors indicate different random MOMDPs.

Memory efficiency To assess memory efficiency, we compared the number of model parameters used in each weight update. As in the original implementation, Park et al. [37] employs 20 copies of the Q-network, resulting in 274,084 parameters per update in the Base-4 and Asym-4 environments (4-dimensional reward), and 274,096 parameters in Asym-16 (16-dimensional reward). In contrast, our method uses only 13,704 and 14,484 parameters, respectively—corresponding to reductions of approximately 95% and 94.7%. Despite this drastic reduction, our method achieves superior max-min performance, as shown in Table 1.

Computational cost efficiency To evaluate the computational complexity of our method, we compared the training wall-time across different environments. Park et al. [37] relies on an extensive soft Q-learning procedure for each weight update, leading to significantly longer training times. In contrast, our single-loop approach updates the weight and policy simultaneously, substantially reducing overall training time. ERAM achieved training time reductions of approximately 67.8% in Base-4, 63.8% in Asym-4, and 68.4% in Asym-16. Similarly, ARAM achieved reductions of approximately 65.5% in Base-4, 63.7% in Asym-4, and 67.6% in Asym-16. Although ARAM includes an additional step for computing correlation vectors, this overhead does not noticeably increase the total runtime. Overall, the computational complexity of ARAM is almost the same as that of ERAM. All experiments were conducted independently on the same hardware to ensure a fair comparison. See Table 2 for a full summary. Note that the memory requirement and computational complexity of ERAM and ARAM are at the same level of GGF-PPO since our w update uses a closed-form solution and the size of w is only the number of objectives.

Envs	ERAM	ARAM	Park et al. [37]	GGF-PPO
Base-4	**111** ± 2.6	120 ± 3.9	346 ± 14	122 ± 4.0
Asym-4	87.2 ± 2.4	87.4 ± 2.4	241 ± 6.3	**84.8 ± 2.2**
Asym-16	**356** ± 27	365 ± 20	1125 ± 95	394 ± 5.5

Table 2: Training wall time (minutes), averaged over five seeds.

6.4 Ablation study

Since $Learner$ of our algorithm is PPO in deep RL cases, ablation study regarding $Learner$ can refer to PPO, and we focus on ablation study of $Adversary$ whose update is given by (13) for ERAM or (40) for ARAM with two hyperparameters λ and τ_w. The algorithms are not so sensitive when β is not near 0 in ERAM or not near 1 in ARAM. When $\beta \approx 0$, ERAM effectively omits the mirror descent term, allowing us to observe the impact of MD. When $\beta \approx 1$, ARAM ignores the adaptive regularizer, highlighting its contribution to performance. Please see Appendix M.3 and Figure 8 for more.

7 Conclusion

In this paper, we have considered MORL with max-min criterion. Exploiting the special max-min and min-max equivalence in this problem, we have reformulated max-min MORL as a two-player zero-sum regularized continuous game. We have proven the existence of a NE of this game, which yields a max-min MORL policy. Then, we have proposed efficient algorithms to find a NE of this game, where the learner can use conventional PPO to update its strategy and the adversary updates its strategy based on a closed-form formula. We have proven last-iterate convergence of the algorithms in the tabular case and demonstrated that the proposed algorithms significantly outperform existing max-min MORL methods. The proposed max-min MORL algorithm has the complexity of PPO and can be used practically for many real-world resource allocation problems.

8 Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025- 00557589, Generative Model Based Efficient Reinforcement Learning Algorithms for Multi-modal Expansion in Generalized Environments) and by the Ministry of Innovation, Science & Technology, Israel grant 1001556423 and ISF grant 2197/22.

References

[1] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *Forty-first International Conference on Machine Learning*, 2024.

[2] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[3] Shubham Aggarwal, Melih Bastopcu, Tamer Başar, et al. Policy optimization finds nash equilibrium in regularized general-sum lq games. *63rd IEEE Conference on Decision and Control, CDC 2024, Milan, Italy, December 16-19, 2024*, pages 3384–3389, 2024.

[4] Lucas N. Alegre. SUMO-RL. `https://github.com/LucasAlegre/sumo-rl`, 2019.

[5] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194. Springer, 2016.

[6] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[7] Yu Bai, Chi Jin, Song Mei, Ziang Song, and Tiancheng Yu. Efficient phi-regret minimization in extensive-form games via online mirror descent. *Advances in Neural Information Processing Systems*, 35:22313–22325, 2022.

[8] Toygun Basaklar, Suat Gumussoy, and Ümit Y. Ogras. Pd-morl: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] Miguel Calvo-Fullana, Santiago Paternain, Luiz FO Chamon, and Alejandro Ribeiro. State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards. *IEEE Transactions on Automatic Control*, 69(7):4275–4290, 2023.

[11] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

[12] Shicong Cen, Yuejie Chi, Simon S Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum markov games. *arXiv preprint arXiv:2210.01050*, 2022.

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

[14] Cyrus Cousins, Kavosh Asadi, Elita Lobo, and Michael Littman. On welfare-centric fair reinforcement learning. In *Reinforcement Learning Conference*, 2024.

[15] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

[16] Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

[17] Eric Eaton, Marcel Hussing, Michael Kearns, Aaron Roth, Sikata Bela Sengupta, and Jessica Sorrell. Intersectional fairness in reinforcement learning with large state and constraint spaces. *arXiv preprint arXiv:2502.11828*, 2025.

[18] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[19] Ziming Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 1991–1999, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

[20] Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[21] Seungyul Han and Youngchul Sung. A max-min entropy framework for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

[22] 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. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

[23] Xiangkun He and Chen Lv. Toward personalized decision making for autonomous vehicles: a constrained multi-objective reinforcement learning technique. *Transportation Research Part C: Emerging Technologies*, 156:104352, 2023.

[24] Yuanzhi He, Biao Sheng, Hao Yin, Di Yan, and Yingchao Zhang. Multi-objective deep reinforcement learning based time-frequency resource allocation for multi-beam satellite communications. *China Communications*, 19(1):77–91, 2022.

[25] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[26] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

[27] Changjian Li and Krzysztof Czarnecki. Urban driving with multi-objective deep reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

[28] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33:12861–12872, 2020.

[29] Peng Li, Song Guo, and Zixue Cheng. Max-min lifetime optimization for cooperative communications in multi-channel wireless networks. *IEEE Transactions on Parallel and Distributed Systems*, 25(6):1533–1542, 2013.

[30] Zhishuai Liu, Weixin Wang, and Pan Xu. Upper and lower bounds for distributionally robust off-dynamics reinforcement learning. *arXiv preprint arXiv:2409.20521*, 2024.

[31] Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2022.

[32] Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. *Advances in neural information processing systems*, 32, 2019.

[33] Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. *arXiv preprint arXiv:2402.15776*, 2024.

[34] Dritan Nace and Michal Pióro. Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial. *IEEE Communications Surveys & Tutorials*, 10(4):5–17, 2008.

[35] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[36] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[37] Giseung Park, Woohyeon Byeon, Seongmin Kim, Elad Havakuk, Amir Leshem, and Youngchul Sung. The max-min formulation of multi-objective reinforcement learning: From theory to a model-free algorithm. *Forty-first International Conference on Machine Learning*, 2024.

[38] Nikolas Patris and Ioannis Panageas. Learning nash equilibria in rank-1 games. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=8utTlmhw8v`.

[39] Nianli Peng and Brandon Fain. Nonlinear multi-objective reinforcement learning with provable guarantees. *arXiv preprint arXiv:2311.02544*, 2023.

[40] Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

[41] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[42] Majid Raeis and Alberto Leon-Garcia. A deep reinforcement learning approach for fair traffic signal control. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2512–2518. IEEE, 2021.

[43] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

[44] Kevin Regan and Craig Boutilier. Robust policy computation in reward-uncertain mdps using nondominated policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1127–1133, 2010.

[45] R Tyrrell Rockafellar. *Convex Analysis*, volume 11. Princeton University Press, 1997.

[46] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48: 67–113, 2013.

[47] Abusayeed Saifullah, David Ferry, Jing Li, Kunal Agrawal, Chenyang Lu, and Christopher D. Gill. Parallel real-time scheduling of dags. *IEEE Trans. Parallel Distributed Syst.*, 25(12): 3242–3252, 2014. doi: 10.1109/TPDS.2013.2297919.

[48] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

[49] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[50] Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.

[51] Umer Siddique, Abhinav Sinha, and Yongcan Cao. Fairness in preference-based reinforcement learning. *arXiv preprint arXiv:2306.09995*, 2023.

[52] Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

[53] Zhuoqing Song, Jason D. Lee, and Zhuoran Yang. Can we find nash equilibria at a linear rate in markov games? In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=eQzLwwGyQrb`.

[54] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[55] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

[56] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

[57] Kankan Wang, Xu Jiang, Nan Guan, Di Liu, Weichen Liu, and Qingxu Deng. Real-time scheduling of DAG tasks with arbitrary deadlines. *ACM Trans. Design Autom. Electr. Syst.*, 24 (6):66:1–66:22, 2019. doi: 10.1145/3358603.

[58] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

[59] Andre Wibisono, Molei Tao, and Georgios Piliouras. Alternating mirror descent for constrained min-max games. *Advances in Neural Information Processing Systems*, 35:35201–35212, 2022.

[60] Yuanyuan Xu, Kun Zhu, Hu Xu, and Jiequ Ji. Deep reinforcement learning for multi-objective resource allocation in multi-platoon cooperative vehicular networks. *IEEE Transactions on Wireless Communications*, 22(9):6185–6198, 2023.

[61] 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.

[62] Tong Yang, Shicong Cen, Yuting Wei, Yuxin Chen, and Yuejie Chi. Federated natural policy gradient methods for multi-task reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

[63] Ephraim Zehavi, Amir Leshem, Ronny Levanda, and Zhu Han. Weighted max-min resource allocation for frequency selective channels. *IEEE Transactions on Signal Processing*, 61(15): 3723–3732, 2013.

[64] Sihan Zeng, Thinh Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum markov games. *Advances in Neural Information Processing Systems*, 35: 34546–34558, 2022.

[65] Wenhao Zhan, Jason D. Lee, and Zhuoran Yang. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. In *The Eleventh International Conference on Learning Representations*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction clearly reflect the theoretical contributions and experimental findings presented in the paper.

   Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix O.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the assumptions and proofs for Theorem 3.1, Theorem 4.1, Theorem 4.3, Corollary 4.2, and Corollary 4.4 in Appendix F.2, Appendix G, Appendix H, Appendix I, and Appendix J.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental environments are described in Appendices L, M, and N, and the hyperparameters are provided in Appendix M.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the source code for the traffic signal control experiment, which constitutes the main experimental result of the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setups and environments are described in Appendix L and M. The source code is provided and hyperparameters are listed in Appendix M.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We used five random seeds for the deep RL benchmarks in Section 6.2 and Appendix N, and fifty random seeds for the tabular settings in Section 6.1 and Appendix L. Figure 1 and Table 2 report the standard deviation across seeds, with shaded areas in the figure and numerical values in the table. Figure 3 shows results from three random seeds for improved visual clarity.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Device specifications are provided in Appendix M.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the social impacts in Appendix O.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not pose risks of misuse requiring special safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use Stable Baselines3 (MIT license), SUMO-RL (MIT license), and MO-Gymnasium (MIT license), all of which are properly cited and used in compliance with their respective licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or human-subject research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects and therefore does not require IRB approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The research does not involve LLMs as an important or original component of the core methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Glossary

Notations	Descriptions
MOMDP	
S	State space
A	Action space
P	Transition dynamics
μ	Initial state distribution
$\mathbf{r}$	Multi-objective reward in the MOMDP
K	number of objectives, i.e., dimension of multi-objective reward
r_k	k-th coordinate of vector reward $\mathbf{r}$, $k = 1, \ldots, K$
γ	Discount factor
π, Π	Policy, Policy space
θ, Θ	Policy parameter, Policy parameter space
w	weight, action of $Adversary$
Values	
$\mathbf{V}^\pi$	$\mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t)\right] \in \mathbb{R}^K$, Value vector under policy π in the MOMDP
V_k^π	$\mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_k(s_t, a_t)\right] \in \mathbb{R}$, k-th coordinate of value vector under policy π in the MOMDP, $k = 1, \ldots, K$
V_w^π	$\langle w, \mathbf{V}^\pi \rangle = \sum_{k=1}^{K} w(k) V_k^\pi$, the weighted value for any $w \in \Delta^K$.
$\mathbf{V}_\tau^\pi$	$\mathbb{E}_{\mu,\pi}\left[\sum_t \gamma^t (\mathbf{r}(s_t, a_t) - \tau \log \pi(a_t \mid s_t) \mathbf{1}_K)\right] \in \mathbb{R}^K$, soft value vector with K objectives and entropy coefficient τ
$V_{k,\tau}^\pi$	k-th coordinate of soft value vector
$V_{w,\tau}^\pi$	$\langle w, \mathbf{V}_\tau^\pi \rangle = \mathbb{E}_{\mu,\pi}\left[\sum_t \gamma^t (\langle w, \mathbf{r}(s_t, a_t) \rangle - \tau \log \pi(a_t \mid s_t))\right] \in \mathbb{R}$, soft value with scalar reward $\langle w, \mathbf{r} \rangle$
$V_{w,\tau}^*, Q_{w,\tau}^*$	Soft optimal values with scalar reward $\langle w, \mathbf{r} \rangle$
Constants	
τ	Entropy coefficient for NPG
τ_w	Entropy coefficient for mirror descent on w
η	Step size for NPG
λ	Step size for mirror descent on w
α	$1 - \frac{\eta\tau}{1-\gamma}$
β	$\frac{1}{\lambda\tau_w + 1}$
Mathematical Notations	
Δ^K	$(K-1)$-Simplex, i.e., $\{w \in \mathbb{R}^K \mid \sum_{k=1}^{K} w(k) = 1, w(k) \geq 0,\ k = 1, \ldots, K\}$
$\langle x, y \rangle$	$\sum_{i=1}^{d} x_i y_i$, inner product for any $x, y \in \mathbb{R}^d$.
$H(w)$	$-\sum_{k=1}^{K} w(k) \log w(k)$, Shannon entropy for any $w \in \Delta^k$
$\tilde{H}(\pi)$	$\mathbb{E}_{\mu,\pi}\left[-\sum_t \gamma^t \log \pi(a_t \mid s_t)\right]$
$\mathrm{softmax}(x)$	$\left(\frac{e^{x_1}}{\sum_{i=1}^{d} e^{x_i}}, \ldots, \frac{e^{x_d}}{\sum_{i=1}^{d} e^{x_i}}\right) \in \mathbb{R}^d$ for any $x \in \mathbb{R}^d$
$\mathbf{1}_d$	all-one-vector with dimension d
D_{KL}	Kullback Leibler divergence
$\lVert \cdot \rVert_1$	l_1-norm, i.e., $\lVert x \rVert_1 = \sum_{i=1}^{d} \lvert x_i \rvert$ for $x \in \mathbb{R}^d$
$\lVert \cdot \rVert_\infty$	l_∞-norm, i.e., $\lVert x \rVert_\infty = \max_i \lvert x_i \rvert$

Table 3: Notations in the main paper

B Relationship between Max-Min Fairness and Other Fairness Criteria in MORL

We can consider three major scalarization methods for MORL, as follows:

- Sum return maximization
$$\max_{\pi} \sum_{i=1}^{K} w_i V_i \tag{21}$$
- Max-min optimization
$$\max_{\pi} \min\{w_1 V_1, \cdots, w_K V_K\} \tag{22}$$
- Proportionally-fair (PF) optimization
$$\max_{\pi} \sum_{i=1}^{K} w_i \log V_i \tag{23}$$

Here, $w_1, \cdots, w_K$ are some given weighting factors. We consider the weighted version of all three methods above for full generality.

The sum return maximization for given weights is a simple single-objective RL for which there are many existing algorithms. Hence, the problem is rather simple.

Consider the PF problem. Since the logarithm function is differentiable, the policy gradient of the PF objective (23) can easily be obtained. That is, let us use $V_i^{\pi_\theta}$ instead of V_i to show the dependence of return on the policy parameter θ explicitly. Then,

$$\frac{\partial}{\partial\theta}\sum_i w_i \log V_i^{\pi_\theta} \quad = \quad \sum_i w_i \frac{1}{V_i^{\pi_\theta}}\frac{\partial V_i^{\pi_\theta}}{\partial\theta}. \tag{24}$$

Note that $\frac{\partial V_i^{\pi_\theta}}{\partial\theta}$ is nothing but the conventional policy gradient, which is given by $\nabla_\theta \log \pi_\theta(s,a)Q_i(s,a)$ [54]. Hence, it is rather straightforward to solve the PF problem for MORL and we only need to track individual Q functions.

In contrast, the max-min problem does not allow such direct differentiation and becomes more complicated. Hence, a specialized method is needed as in [37] or approximate approaches were considered as in [19, 39, 50]. In this paper, we solve the exact max-min MORL problem and provide a very efficient algorithm for this problem.

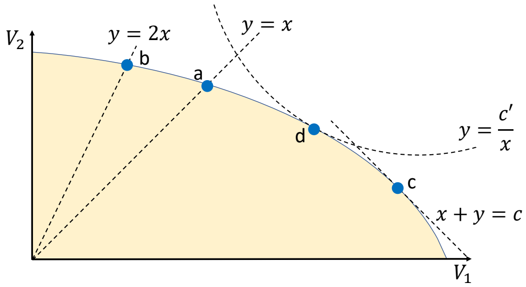


Figure 4: Relationship between max-min fairness and other fairness notions in MORL.

Figure 4 illustrates the solution of each method for equal weights with two objectives. In the figure, all of points **a**, **b**, **c**, and **d** are Pareto optimal. The max-min solution, point **a**, corresponds to the Pareto-optimal point that achieves Pareto efficiency with $V_1 = V_2$ due to equalizer rule [63]. Note that by weighting w_1 and w_2, we can achieve the Pareto-boundary point on which $w_1V_1 = w_2V_2$ due to the same equalizer rule, e.g., Point **b** in the figure.

Point **c** is the sum return maximization point on which the Pareto-boundary is tangential with the straight line $V_1 + V_2 = c$ for some c. Again, we can change the slope of the line by considering weights as $w_1V_1 + w_2V_2 = c$.

Point **d** is the PF point on which Pareto-boundary is tangential to the curve $V_1V_2 = c'$ for some c' because $\sum_i \log V_i = \log \prod_i V_i$. Note that the PF point is a compromised point between the sum maximization point and the max-min point.

C Brief Summary of the Prior Work: Convex Optimization Formulation of Max-Min MORL [37]

Based on the fact that value function is linear in state-action occupancy measure [54, 41], Park et al. [37] rewrote the max-min MORL without entropy regularization in terms of state-action occupancy measure (**P0** in Park et al. [37]).

$$\mathbf{P0}: \quad \max_{d} \min_{k=1,\ldots,K} \sum_{s,a} d(s,a) r_k(s,a) \tag{25}$$

$$\sum_{a'} d(s',a') = \mu(s') + \gamma \sum_{s,a} P(s'|s,a)d(s,a), \;\; \forall s' \tag{26}$$

$$d(s,a) \geq 0, \;\; \forall s, a. \tag{27}$$

By taking dual problem and linear programming formulation of value iteration [41], Park et al. [37] has shown that solving **P0** is equivalent to solving the following convex optimization problem **P2**. The convexity of the objective function in **P2**, and the equivalence of **P0** and **P2** are addressed in Theorem 3.1 and Theorem 3.2 in Park et al. [37], respectively.

$$\textbf{P2} \text{ in Park et al. [37] : } \min_{w\in\Delta^K} \sum_s \mu(s) V_w^*(s) \tag{28}$$

where $V_w^*(\cdot)$ is optimal value function with linearly scalarized reward $\langle w, \mathbf{r}(s,a)\rangle$.

However, this value iteration based method can fail to find optimal policy in max-min MORL even in a simple one-state MOMDP [46, 37]. This issue is called indeterminacy; that is, there exists an MOMDP which has no max-min optimal policy, which is stationary and deterministic. To circumvent this issue, [37] adopted entropy regularization to ensure the policy stochastic and established convex programming reformulation similar to above:

$$\textbf{P0':} \max_d \min_{k=1,\ldots,K} \sum_{s,a} d(s,a)\left\{r_k(s,a) + \tau H(\pi^d(\cdot|s))\right\} \tag{29}$$

$$\begin{aligned} \sum_{a'} d(s',a') =& \mu(s') + \gamma \sum_{s,a} P(s'|s,a) d(s,a), \ \ \forall s' & (30)\\ d(s,a) \geq& 0, \ \ \forall s, a & (31)\end{aligned}$$

where $\pi^d(a|s) := \frac{d(s,a)}{\sum_{a'} d(s,a')}$.

Similar to above, solving **P0'** is equivalent to solving the following convex optimization problem **P2'**. The convexity of the objective function in **P2'**, and the equivalence of **P0'** and **P2'** are addressed in Theorem 4.1 and Theorem 4.2 in Park et al. [37], respectively.

$$\textbf{P2'} \text{ in Park et al. [37] : } \min_{w\in\Delta^K} \sum_s \mu(s) V_{w,\tau}^*(s) \tag{32}$$

where $V_{w,\tau}^*(\cdot)$ is the soft optimal value function with scalar reward $\langle w, \mathbf{r}(s,a)\rangle$ which is the solution of soft Bellman equation in v.

$$v(s) = \tau \log \sum_a \exp[\frac{1}{\tau}\{\langle w, \mathbf{r}(s,a)\rangle + \gamma \sum_{s'} P(s'|s,a) v(s')\}], \ \forall s$$

D More on Adaptive Regularization of $Adversary$

With the adaptive regularizer $-D_{KL}(w\|c_t)$, where

$$c_t := c(\pi_t) = \text{softmax}\left(\mathbb{E}_{(s,a)\sim d^{\pi_t}}\left[r_k(s,a)\cdot r_{i'_t}(s,a)\right]\right)_{k=1}^K, \tag{33}$$

and where i'_t denotes the index of the worst-performing objective at iteration t, i.e., $V_{i'_t}^{\pi_t} \leq V_k^{\pi_t}$ for all $k = 1, \ldots, K$, the $Adversary$'s update rule in Equation (12) becomes:

$$\begin{aligned} w_{t+1} =& \operatorname*{arg\,max}_{w\in\Delta^K} \ \lambda\langle \nabla_w(-\langle w, \mathbf{V}^{\pi_{\theta_t}}\rangle)|_{w=w_t}, \ w\rangle - \lambda\tau_w D_{KL}(w\|c_t) - D_{KL}(w\|w_t) & (34)\\ =& \operatorname*{arg\,max}_{w\in\Delta^K} \langle -\lambda \mathbf{V}^{\pi_{\theta_t}} + \lambda\tau_w \log c_t + \log w_t, \ w\rangle + (\lambda\tau_w + 1) H(w) & (35)\\ =& \operatorname*{arg\,max}_{w\in\Delta^K} \langle -\frac{1-\beta}{\tau_w}\mathbf{V}^{\pi_{\theta_t}} + (1-\beta)\log c_t + \beta \log w_t, \ w\rangle + H(w). & (36)\end{aligned}$$

This leads to the following closed-form solution [9]:

$$w_{t+1} = \text{softmax}\left(-\frac{1-\beta}{\tau_w}\mathbf{V}^{\pi_t} + \beta\log w_t + (1-\beta)\log c(\pi_t)\right). \tag{37}$$

The pseudo-code of the deep RL implementation of ARAM combined with PPO is provided in Section E.3.

E Pseudo Codes

E.1 ERAM

Algorithm 1 Adversary with regularizer for max-min MORL employing exact policy evaluation

```
Input: initial policy parameter $\theta_0$, initial weight $w_0$, number of iterations $T$, NPG step size $\eta$, MD step size $\lambda$, regularizer coefficients: $\tau, \tau_w$
for $t = 1$ to $T$ do
    Evaluate current policy $\pi_t$ and obtain $\mathbf{V}^{\pi_t}$
    $\theta_{t+1} \leftarrow \theta_t + \eta F^{\dagger}(\theta_t) \sum_{k=1}^{K} w_t(k) \nabla_\theta V_{k,\tau}^{\pi_{\theta_t}}$    ▷ update $Learner$'s strategy
    $w_{t+1} \leftarrow \mathrm{softmax}\left(-\frac{1-\beta}{\tau_w}\mathbf{V}^{\pi_t} + \beta \log w_t\right)$    ▷ update $Adversary$'s strategy in closed form
end for
Return: $\pi_T, w_T$
```

In the tabular case with softmax policy, the θ update is simplified as

$$\pi_{\theta_{t+1}}(a|s) \propto (\pi_{\theta_t}(a|s))^{\alpha} \exp(\frac{1-\alpha}{\tau} Q_{\tau}^{\pi_{\theta_t}}(s,a)). \tag{38}$$

In the deep learning setting, the θ update can be replaced with PPO [49], and the corresponding θ update is seen in Appendix E.3.

E.2 ERAM with Approximate Policy Evaluation

Algorithm 2 ERAM employing approximate policy evaluation

```
Input: initial policy parameter $\theta_0$, initial weight $w_0$, number of iterations $T$, NPG step size $\eta$, MD step size $\lambda$, regularizer coefficients: $\tau, \tau_w$
for $t = 1$ to $T$ do
    Obtain approximate policy evaluation $\widehat{\mathbf{V}}^{\pi_t}$ for current policy $\pi_t$
    $\theta_{t+1} \leftarrow \theta_t + \eta F^{\dagger}(\theta_t) \sum_{k=1}^{K} w_t(k) \nabla_\theta V_{k,\tau}^{\pi_{\theta_t}}$    ▷ update $Learner$'s strategy with $\widehat{\mathbf{V}}^{\pi_t}$, e.g., (39)
    $w_{t+1} \leftarrow softmax(-\frac{1-\beta}{\tau_w}\widehat{\mathbf{V}}^{\pi_t} + \beta \log w_t)$    ▷ update $Adversary$'s strategy
end for
Return: $\pi_T, w_T$
```

In the tabular case with softmax policy, the θ update with approximate policy evaluation is simplified as

$$\pi_{\theta_{t+1}}(a|s) \propto (\pi_{\theta_t}(a|s))^{\alpha} \exp(\frac{1-\alpha}{\tau} \widehat{Q}_{\tau}^{\pi_{\theta_t}}(s,a)). \tag{39}$$

In the deep learning setting, the θ update can be replaced with PPO [49], and the corresponding θ update is seen in Appendix E.3.

E.3 ARAM with PPO for the Learner Update

Algorithm 3 Adversary with adaptive regularizer for max-min MORL with multi-objective variant of PPO

```
Input: policy network $\pi_\theta$ and state-value vector network $\mathbf{V}_\phi$, initial parameters $\theta_0, \phi_0$, clip coefficient $\epsilon_{clip}$, initial weight $w_0$, number of iterations $T$, MD step size $\lambda$, regularizaion coefficients: $\tau_{PPO}, \tau_w$
for $t = 1$ to $T$ do
    Collect samples from $\pi_{\theta_t}$ and update critic buffer $\mathcal{B}_{critic}$
    $\widehat{\mathbf{V}}_t := \mathbf{V}_{\phi_t}$        ▷ Get estimate of value for current policy with current critic network
    Compute $c_t$:
        $$c(\pi_t) = \text{softmax}\left(\mathbb{E}_{(s,a)\sim d^{\pi_{\theta t}}}\left[r_k(s,a) r_{i'_t}(s,a)\right]\right)_{k=1}^{K}$$
    Update $Learner$'s strategy with PPO
        Optimize $L_{clip}(\theta; \theta_t, w_t)$ to obtain $\theta_{t+1}$
        Optimize $L_{critic}(\phi)$ to obtain $\phi_{t+1}$
    Update $Adversary$'s strategy
        $$w_{t+1} \leftarrow \text{softmax}\left(-\frac{1-\beta}{\tau_w}\widehat{\mathbf{V}}_t + \beta \log w_t + (1-\beta)\log c_t\right) \quad (40)$$
        ▷
end for
Return: $\pi_T, w_T$
```

Here, $L_{clip}(\theta; \theta_t, w_t) = -\mathbb{E}_{(s,a)\sim d^{\pi_{\theta_t}}}[\min\{r_t(\theta)(s,a)\langle w_t, \widehat{\mathbf{A}}(s,a)\rangle, clip(r_t(\theta)(s,a), 1-\epsilon_{clip}, 1+\epsilon_{clip})\langle w_t, \widehat{\mathbf{A}}(s,a)\rangle\} + \tau_{PPO} H(\pi_\theta(\cdot|s))]$ with estimated vector advantage $\widehat{\mathbf{A}}(s,a) \in \mathbb{R}^K$, and $L_{critic}(\phi) = \mathbb{E}_{(s,\mathbf{V}_{s,target})\sim\mathcal{B}_{critic}}[\|\mathbf{V}_\phi(s) - \mathbf{V}_{s,target}\|^2]$ with the estimated return vector at s replacing $\mathbf{V}_{s,target}$.

The closed-form update formula for $Adversary$'s strategy w with the $D(w||c_t)$ regularizer is shown above.

F Supplementary Materials for Section 3

F.1 Existence of a saddle point in (6)

Proposition F.1. *If* $\max_\pi \min_{w\in\Delta^K}\langle w, \mathbf{V}^\pi_\tau\rangle = \min_{w\in\Delta^K}\max_\pi\langle w, \mathbf{V}^\pi_\tau\rangle$ *holds, then a saddle point of* $\langle w, \mathbf{V}^\pi_\tau\rangle$ *exists.*

Proof. By assumption, $\max_\pi \min_{w\in\Delta^K}\langle w, V^\pi_\tau\rangle = \min_{w\in\Delta^K}\max_\pi\langle w, V^\pi_\tau\rangle$ -(1) holds. By extreme value theorem, $w^* := \arg\min_{w\in\Delta^K}\max_\pi\langle w, \mathbf{V}^\pi_\tau\rangle$ -(2) and $\pi^* := \arg\max_\pi \min_{w\in\Delta^K}\langle w, \mathbf{V}^\pi_\tau\rangle$ -(3) exist. (Note that the policy space $\Pi = \Delta(A)^{|S|}$ is compact due to Tychonoff's theorem.) Then,

$$\begin{aligned}
\langle w^*, \mathbf{V}^{\pi^*}_\tau\rangle &\geq \min_w \langle w, \mathbf{V}^{\pi^*}_\tau\rangle && (41)\\
&= \max_\pi \min_w \langle w, \mathbf{V}^\pi_\tau\rangle \ (\because\ (3)) && (42)\\
&= \min_w \max_\pi \langle w, \mathbf{V}^\pi_\tau\rangle \ (\because\ (1)) && (43)\\
&= \max_\pi \langle w^*, \mathbf{V}^\pi_\tau\rangle \ (\because\ (2)) && (44)\\
&\geq \langle w^*, \mathbf{V}^{\pi^*}_\tau\rangle. && (45)
\end{aligned}$$

Therefore,

$$\langle w^*, \mathbf{V}^\pi_\tau\rangle \leq \max_\pi \langle w^*, \mathbf{V}^\pi_\tau\rangle = \langle w^*, \mathbf{V}^{\pi^*}_\tau\rangle = \min_w \langle w, \mathbf{V}^{\pi^*}_\tau\rangle \leq \langle w, \mathbf{V}^{\pi^*}_\tau\rangle,\ \forall w, \pi. \quad (46)$$

This implies that (w^*, π^*) is a saddle point of $\langle w, \mathbf{V}^\pi_\tau\rangle$, which concludes the existence of a saddle point of $\langle w, \mathbf{V}^\pi_\tau\rangle$. □

F.2 Proof for Theorem 3.1

Proof. Let $(\bar{\pi}, \bar{w})$ be a saddle point in (6) and for any $w \in \Delta^K$, define $\pi^*(w) := argmax_\pi \langle w, \mathbf{V}_\tau^\pi \rangle$. Let w^* be the optimal solution of minimization reformulation of entropy-regularized max-min MORL (3).

By Theorem 4.2 in Park et al. [37], if w is any optimal solution of (3), whose objective function is $V_{\cdot,\tau}^* \triangleq V_{\cdot,\tau}^{\pi^*(\cdot)} = \langle \cdot, V_\tau^{\pi^*(\cdot)} \rangle$ with the domain Δ^K, the policy defined by $\pi_w(a|s) := \exp\left(\frac{Q_{w,\tau}^*(s,a) - V_{w,\tau}^*(s)}{\tau}\right) \triangleq \exp\left(\frac{Q_{w,\tau}^{\pi^*(w)}(s,a) - V_{w,\tau}^{\pi^*(w)}(s)}{\tau}\right)$ is an optimal solution of entropy-regularized max-min MORL (1) where $Q_{w,\tau}^*$ and $V_{w,\tau}^*$ are soft optimal values for scalar reward $\langle w, \mathbf{r} \rangle$. Therefore, we show that $\bar{w}$ is an optimal solution of (3). This concludes that the induced policy $\pi_{\bar{w}}(a|s) := \exp\left(\frac{Q_{\bar{w},\tau}^*(s,a) - V_{\bar{w},\tau}^*(s)}{\tau}\right)$ is an optimal solution of entropy-regularized max-min MORL (1).

By the definition of saddle point,

$$\langle \bar{w}, \mathbf{V}_\tau^\pi \rangle \leq \langle \bar{w}, \mathbf{V}_\tau^{\bar{\pi}} \rangle \leq \langle w, \mathbf{V}_\tau^{\bar{\pi}} \rangle, \ \ \forall w, \pi. \tag{47}$$

In particular, $\bar{\pi} = \pi^*(\bar{w})$ by (47), i.e. $\bar{\pi}$ is soft optimal policy for scalar reward $\langle \bar{w}, \mathbf{r} \rangle$. Then,

$$\begin{align}
\langle \bar{w}, \mathbf{V}_\tau^{\bar{\pi}} \rangle &= \min_{w'} \langle w', \mathbf{V}_\tau^{\bar{\pi}} \rangle \ (\because (47)) \tag{48}\\
&= \min_{w'} \langle w', \mathbf{V}_\tau^{\pi^*(\bar{w})} \rangle \ (\because \bar{\pi} = \pi^*(\bar{w})) \tag{49}\\
&\leq \langle w^*, \mathbf{V}_\tau^{\pi^*(\bar{w})} \rangle \tag{50}\\
&\leq \langle w^*, \mathbf{V}_\tau^{\pi^*(w^*)} \rangle \tag{51}\\
&\leq \langle \bar{w}, \mathbf{V}_\tau^{\pi^*(\bar{w})} \rangle \tag{52}\\
&\quad (\because w^* \text{ is an optimal solution of (3) with objective function } V_{\cdot,\tau}^* \triangleq V_{\cdot,\tau}^{\pi^*(\cdot)} = \langle \cdot, V_\tau^{\pi^*(\cdot)} \rangle\ ) \tag{53}\\
&= \langle \bar{w}, \mathbf{V}_\tau^{\bar{\pi}} \rangle \ (\because \bar{\pi} = \pi^*(\bar{w})). \tag{54}
\end{align}$$

Therefore, $\langle \bar{w}, \mathbf{V}_\tau^{\bar{\pi}} \rangle = \langle w^*, \mathbf{V}_\tau^{\pi^*(w^*)} \rangle$, which means that $\bar{w}$ is an optimal solution of (3). Then, for any s, a,

$$\begin{align}
\pi_{\bar{w}}(a|s) &= \exp\left(\frac{Q_{\bar{w},\tau}^*(s,a) - V_{\bar{w},\tau}^*(s)}{\tau}\right) \tag{55}\\
&= \exp\left(\frac{Q_{\bar{w},\tau}^{\pi^*(\bar{w})}(s,a) - V_{\bar{w},\tau}^{\pi^*(\bar{w})}(s)}{\tau}\right) \ (\because \text{definition of } \pi^*(\cdot)) \tag{56}\\
&= \exp\left(\frac{Q_{\bar{w},\tau}^{\bar{\pi}}(s,a) - V_{\bar{w},\tau}^{\bar{\pi}}(s)}{\tau}\right) \ (\because \bar{\pi} = \pi^*(\bar{w})) \tag{57}\\
&= \bar{\pi}(a|s) \ (\because \bar{\pi} \text{ is soft optimal policy for scalar reward } \langle \bar{w}, \mathbf{r} \rangle). \tag{58}
\end{align}$$

Therefore, $\bar{\pi}$ is an optimal solution of entropy-regularized max-min MORL (1). □

F.3 Remark for replacing MD with NPG

Remark F.2. *The MD update rule (13) can be written as*

$$w_{t+1}(k) \propto (w_t(k))^\beta \exp\left(-\frac{1-\beta}{\tau_w} V_k^{\pi_t}\right), \ \ \forall k = 1, \ldots, K. \tag{59}$$

Recall that the NPG update with softmax policy (11) *is*

$$\pi_{\theta_{t+1}}(a|s) \propto (\pi_{\theta_t}(a|s))^\alpha \exp\left(\frac{1-\alpha}{\tau} Q_\tau^{\pi_{\theta_t}}(s,a)\right). \tag{60}$$

We can observe that the update rules for MD and NPG have similar structures under the softmax policy family. This gives an intuition for why replacing MD on θ (9) with NPG does not appear to be absurd.

G Proof of Theorem 4.1

Note that the strategy of $Learner$ is a policy parameter θ_t, it suffice to analyze the convergence of resulting policy $\pi_t := \pi_{\theta_t}$. Similar to [11], we define auxiliary variables that are needed for our proof.

$$\begin{aligned}
\xi_0(s,a) :=& \| \exp(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau)\|_1 \pi_0(a|s) && (61)\\
\xi_{t+1}(s,a) :=& (\xi_t(s,a))^{\alpha} e^{\frac{1-\alpha}{\tau} Q^{\pi_t}_{w_t,\tau}(s,a)} && (62)\\
\kappa_0(k) =& \| \exp(-\mathbf{V}^{\pi^*}_\tau/\tau_w)\|_1 w_0(k) && (63)\\
\kappa_{t+1}(k) =& (\kappa_t(k))^{\beta} e^{-\frac{1-\beta}{\tau_w}\mathbf{V}^{\pi_t}_\tau} && (64)\\
(\pi^*, w^*) :& \text{ Nash equilibrium in } \mathcal{RG} && (65)\\
Q^*_{w,\tau} :=& Q^{\pi^*(w)}_{w,\tau} \triangleq \max_\pi Q^\pi_{w,\tau} \text{ (soft optimal } Q\text{-value w.r.t. reward } \langle w, \mathbf{r}\rangle) && (66)\\
V^*_{w,\tau} :=& V^{\pi^*(w)}_{w,\tau} \triangleq \max_\pi V^\pi_{w,\tau} \text{ (soft optimal } V\text{-value w.r.t. reward } \langle w, \mathbf{r}\rangle) && (67)
\end{aligned}$$

In words, $\xi_t(s,\cdot)$ and κ_t are unnormalized $\pi_t(\cdot|s)$ and w, respectively. That is, $\pi_t(a|s) = \xi_t(s,a)/\|\xi_t(s,\cdot)\|_1,\ \forall t, s, a$ and $w_t(k) = \kappa_t(k)/\|\kappa_t\|_1,\ \forall t, k$.

G.1 Introducing optimality gaps

Optimality gap for π

We note the 1-Lipschitzness of log-sum-exponential (lse) function, which was also mentioned in [11]. Define the function $lse(x) := \log \sum_{i=1}^d e^{x_i}$. For $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
|lse(x) - lse(y)| =& |\langle \nabla lse(z), x - y\rangle| \\
& \text{(where } z = tx + (1-t)y \text{ for some } t \in (0,1)\text{, by mean value theorem)}\\
=& |\langle \text{softmax}(z), x - y\rangle| \\
\le& \|x - y\|_\infty \sum_{i=1}^d \frac{e^{z_i}}{\sum_{j=1}^d e^{z_j}} = \|x-y\|_\infty.
\end{aligned}$$

Then, for each s, a,

$$\begin{aligned}
|\log \pi^*(a|s) - \log \pi_t(a|s)| =& |(Q^{\pi^*}_{w^*,\tau}(s,a)/\tau - V^{\pi^*}_{w^*,\tau}(s)/\tau) - (\log \xi_t(s,a) - \log \|\xi_t(s,\cdot)\|_1)| \\
=& |(Q^{\pi^*}_{w^*,\tau}(s,a)/\tau - \log \xi_t(s,a)) - (V^{\pi^*}_{w^*,\tau}(s)/\tau - \log \|\xi_t(s,\cdot)\|_1)| \\
\le& |Q^{\pi^*}_{w^*,\tau}(s,a)/\tau - \log \xi_t(s,a)| \\
&+ |lse(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau) - lse(\log \xi_t(s,a))| \\
&\quad (\because V^{\pi^*}_{w^*,\tau}(s)/\tau = lse(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau) \text{ holds for soft optimal value}) \\
\le& |Q^{\pi^*}_{w^*,\tau}(s,a)/\tau - \log \xi_t(s,a)| + |Q^{\pi^*}_{w^*,\tau}(s,a)/\tau - \log \xi_t(s,a)| \\
&\quad (\because \text{1-Lipschitzness of } lse) \\
=& 2|Q^{\pi^*}_{w^*,\tau}(s,a)/\tau - \log \xi_t(s,a)|
\end{aligned}$$

where the first equality utilized the fact that $\pi^* = RBR(w^*)$ (regularized best-response), i.e. soft optimal policy with respect to reward $\langle w^*, \mathbf{r}\rangle$, and the property of soft value function and soft optimal policy $\pi^*(a|s) = \exp\left(\frac{Q^{\pi^*}_{w^*,\tau}(s,a) - V^{\pi^*}_{w^*,\tau}(s)}{\tau}\right)$ established in [35].

Therefore,

$$\|\log \pi^* - \log \pi_t\|_\infty \le \frac{2}{\tau}\|Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_t\|_\infty. \tag{68}$$

In words, the term $\|Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_t\|_\infty$ bounds the gap between policy π_t in t-th iteration and Nash policy π^*.

Optimality gap for w

Lemma G.1. $\|w - w'\|_\infty \le \|\log w - \log w'\|_\infty,\ \forall w, w' \in int(\Delta^K)$ *where the logarithm is taken element-wisely and* $int(D)$ *denotes interior of set* D.

Proof. Let $x, y \in int(\Delta^K)$, interior of simplex, and define $f(x) := [\log x_1, \dots, \log x_K]^T$ for any $x \in int(\Delta^K)$. Let $a = f(x)$, $b = f(y)$ and $g(x) := [e^{x_1}, \dots, e^{x_K}]^T$ for any $x \in \mathbb{R}^K$. By the mean value theorem for multi-variate function g,

$$g(b) - g(a) = J_g(c)(b - a)$$

where $J_g(c)$ is Jacobian matrix of g at c and $c = ta + (1-t)b,\ t \in (0,1)$. Specifically, the Jacobian matrix becomes a diagonal matrix $J_g(c) = diag(e^{c_1}, \dots, e^{c_K})$. Then,

$$\begin{aligned}\|g(b) - g(a)\|_\infty = \|J_g(c)(b-a)\|_\infty \le& \|J_g(c)\|_\infty \|b - a\|_\infty \\ =& \max\{e^{c_1}, \dots, e^{c_K}\}\|b - a\|_\infty\end{aligned}$$

Here, since $x_i, y_i < 1$ ($\because x, y \in int(\Delta^K)$), $a_i = \log x_i < 0$ and $b_i = \log y_i < 0$. Hence, $c_i = ta_i + (1-t)b_i < 0$. These results in $\max\{e^{c_1}, \dots, e^{c_K}\} < 1$. By plugging in $a = f(x)$, $b = f(y)$ above, we conclude that $\|y - x\|_\infty < \|\log y - \log y\|_\infty$ for $x, y \in int(\Delta^K)$. □

Due to the lemma G.1, it suffice to analyze $\|\log w^* - \log w_t\|_\infty$ instead of $\|w^* - w_t\|_\infty$.
Note that w^* is the best-response to π^* with respect to $u^{\mathcal{RG}}$ since (π^*, w^*) is Nash equilibrium in $\mathcal{RG}$, and w^* can be written as below.

$$\begin{aligned}w^* =& \operatorname*{arg\,max}_{w \in \Delta^K} -\langle w, \mathbf{V}_\tau^{\pi^*}\rangle + \tau_w H(w) && (69)\\ =& \operatorname{softmax}\left(-\frac{1}{\tau_w}\mathbf{V}_\tau^{\pi^*}\right) && (70)\end{aligned}$$

Using this,

$$\begin{aligned}\log w^*(k) - \log w_t(k) =& \log \frac{\exp\left(-\frac{1}{\tau_w} V_{k,\tau}^{\pi^*}\right)}{\|\exp\left(-\frac{1}{\tau_w}\mathbf{V}_\tau^{\pi^*}\right)\|_1} - \log\frac{\kappa_t(k)}{\|\kappa_t\|_1} && (71)\\ =& -\frac{1}{\tau_w} V_{k,\tau}^{\pi^*} - \log\kappa_t(k) - (lse(-\frac{1}{\tau_w}\mathbf{V}_\tau^{\pi^*}) - lse(\log\kappa_t)) && (72)\end{aligned}$$

By taking absolute value for both side and applying triangular inequality,

$$\begin{aligned}|\log w^*(k) - \log w_t(k)| \le& |-\frac{1}{\tau_w}V_{k,\tau}^{\pi^*} - \log\kappa_t(k)| + |lse(-\frac{1}{\tau_w}\mathbf{V}_\tau^{\pi^*}) - lse(\log\kappa_t)| && (73)\\ \le& |-\frac{1}{\tau_w}V_{k,\tau}^{\pi^*} - \log\kappa_t(k)| + \|-\frac{1}{\tau_w}\mathbf{V}_\tau^{\pi^*} - \log\kappa_t\|_\infty && (74)\\ & (\because \text{1-Lipschitzness of } lse) && (75)\end{aligned}$$

By taking max_k for both sides, obtain the following bound.

$$\|\log w^* - \log w_t\|_\infty \le \frac{2}{\tau_w}\| - \mathbf{V}_\tau^{\pi^*} - \tau_w \log\kappa_t\|_\infty \tag{76}$$

In particular, combining this bound with the lemma G.1, obtain the following bounds.

$$\|w^* - w_t\|_\infty \le \|\log w^* - \log w_t\|_\infty \le \frac{2}{\tau_w}\| - \mathbf{V}_\tau^{\pi^*} - \tau_w\log\kappa_t\|_\infty \tag{77}$$

Optimality gap for soft Q
The optimality gap for soft Q at iteration t can be expressed as follows.

$$\|Q_{w^*,\tau}^{\pi^*} - Q_{w_t,\tau}^{\pi_t}\|_\infty \tag{78}$$

Supplementary term
Following [11], we use the following supplementary term which helps to analyze the optimality gap for soft Q.

$$\max\{0, -\min_{s,a}(Q_{w_t,\tau}^{\pi_t} - \tau\log\xi_t)\} \tag{79}$$

G.2 Recursive bounds for optimality gaps and supplementary term

We establish recursive bounds for the following three optimality gaps and the supplementary term: $\|Q^{\pi^*}_{w^*,\tau} - \tau\log\xi_t\|_\infty$ (optimality gap for policy (68)), $\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty$ (optimality gap for soft Q function), $\|-\mathbf{V}^{\pi^*}_\tau - \tau_w\log\kappa_t\|_\infty$ (optimality gap for w (77)) and $\max\{0, -\min_{s,a}(Q^{\pi_t}_{w_t,\tau} - \tau\log\xi_t)\}$ (supplementary term, used for optimality gap for soft Q function). After then, following [11], we establish linear system with these optimality gaps to show convergence.

Recursive bounds: optimality gap for policy For each (s,a),

$$Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\log\xi_{t+1}(s,a) = Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\alpha\log\xi_t(s,a) - (1-\alpha)Q^{\pi_t}_{w_t,\tau}(s,a) \ (\because (62)) \tag{80}$$

$$= \alpha(Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\log\xi_t(s,a)) \tag{81}$$

$$+ (1-\alpha)(Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a)) \tag{82}$$

Thus,

$$\|Q^{\pi^*}_{w^*,\tau} - \tau\log\xi_{t+1}\|_\infty \le \alpha\|Q^{\pi^*}_{w^*,\tau} - \tau\log\xi_t\|_\infty + (1-\alpha)\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty \tag{83}$$

Recursive bounds: optimality gap for soft Q function Since π (thus, θ) is a max-player and w is a min-player in our two-player zero-sum regularized game, we cannot guarantee monotonicity of $Q^{\pi_t}_{w_t,\tau}(s,a)$ in t. Thus, we establish both upper and lower bounds for $Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a)$ and these will offer an upper bound for $|Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a)|$, thus upper bound for $\|Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a)\|_\infty$.

1) Upper bound
For each s, a,

$$Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) \tag{84}$$

$$=\langle w^*, \mathbf{r}(s,a)\rangle + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^{\pi^*}_{w^*,\tau}(s')] - \langle w_{t+1}, r(s,a)\rangle - \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^{\pi_{t+1}}_{w_{t+1},\tau}(s')] \tag{85}$$

$$=\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*_{w^*,\tau}(s')] \tag{86}$$

$$-\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\mathbb{E}_{a'\sim\pi_{t+1}(\cdot|s')}[Q^{\pi_{t+1}}_{w_{t+1},\tau}(s',a') - \tau\log\pi_{t+1}(s',a')] \tag{87}$$

$$=\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\tau\log\|e^{Q^*_{w^*,\tau}(s',\cdot)/\tau}\|_1] \tag{88}$$

$$-\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\mathbb{E}_{a'\sim\pi_{t+1}(\cdot|s')}\left[Q^{\pi_{t+1}}_{w_{t+1},\tau}(s',a') - \tau\log\frac{\xi_{t+1}(s',a')}{\|\xi_{t+1}(s',\cdot)\|_1}\right] \tag{89}$$

$$=\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle \tag{90}$$

$$+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\tau\log\|e^{Q^{\pi^*}_{w^*,\tau}(s',\cdot)/\tau}\|_1 - \tau\log\|\xi_{t+1}(s',\cdot)\|_1] \quad -(A) \tag{91}$$

$$-\gamma\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_{t+1}(\cdot|s')}[Q^{\pi_{t+1}}_{w_{t+1},\tau}(s',a') - \tau\log\xi_{t+1}(s',a')] \quad -(B) \tag{92}$$

In the third line, the second term used $V^{\pi^*}_{w^*,\tau} = V^*_{w^*,\tau}$ since π^* is the best-response to w^* with respect to $u^{\mathcal{RG}}$ and the third term used soft Bellman equation. In the fourth line, used $V^*_{w^*,\tau}(s') = \tau\log\|e^{Q^*_{w^*,\tau}(s',\cdot)/\tau}\|_1$ which holds for soft optimal value function.

We can easily upper bound the line (90), $\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle$, by $Kr_{max}\|w^* - w_{t+1}\|_\infty$ due to $H\ddot{o}lder$ inequality and $\|\mathbf{r}(s,a)\|_1 \le Kr_{max}\ \forall s,a$ where $r_{max} = \max_{s,a,k}|r_k(s,a)|$. By (77),

$$\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle \tag{93}$$

$$\le Kr_{max}\|w^* - w_{t+1}\|_\infty \tag{94}$$

$$\le \frac{2Kr_{max}}{\tau_w}\|-\mathbf{V}^{\pi^*}_\tau - \tau_w\log\kappa_{t+1}\|_\infty \tag{95}$$

From now, we derive the upper bound for (A).

For each s,

$$\tau \log \|e^{Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau}\|_1 - \tau \log \|\xi_{t+1}(s,\cdot)\|_1 \tag{96}$$
$$=\tau(lse(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau) - lse(\log \xi_{t+1}(s,\cdot))) \tag{97}$$
$$\leq \tau \|Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau - \log \xi_{t+1}(s,\cdot)\|_\infty \ (\because \text{1-Lipschitzness of } lse) \tag{98}$$
$$=\|Q^{\pi^*}_{w^*,\tau}(s,\cdot) - \tau \log \xi_{t+1}(s,\cdot)\|_\infty \tag{99}$$
$$\leq \|Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_{t+1}\|_\infty \tag{100}$$

Therefore,

$$(A) = \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\tau \log \|e^{Q^{\pi^*}_{w^*,\tau}(s',\cdot)/\tau}\|_1 - \tau \log \|\xi_{t+1}(s',\cdot)\|_1] \tag{101}$$
$$\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\|Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_{t+1}\|_\infty] \tag{102}$$
$$= \gamma \|Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_{t+1}\|_\infty \tag{103}$$

From now, we derive the upper bound for (B).

To establish the upper bound, following Cen et al. [11] and Yang et al. [62], we use supplementary term. For each s, a,

$$Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a) \tag{104}$$
$$\geq \min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)\}, \tag{105}$$

then,

$$-(Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)) \tag{106}$$
$$\leq - \min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)\} \tag{107}$$
$$\leq \max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)\}\}. \tag{108}$$

Using the last term, establish the upper bound for (B) as follows.

$$(B) = -\gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_{t+1}(\cdot|s')}[Q^{\pi_{t+1}}_{w_{t+1},\tau}(s',a') - \tau \log \xi_{t+1}(s',a')] \tag{109}$$
$$\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_{t+1}(\cdot|s')}[\max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)\}\}] \tag{110}$$
$$= \gamma \max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)\}\} \tag{111}$$

Combining (95), (103) and (111), obtain the upper bound for optimality gap for soft Q as follows. For each s, a,

$$Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) \tag{112}$$
$$\leq \frac{2Kr_{max}}{\tau_w}\| - \mathbf{V}^{\pi^*}_\tau - \tau_w \log \kappa_{t+1}\|_\infty + \gamma \|Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_{t+1}\|_\infty \tag{113}$$
$$+ \gamma \max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)\}\} \tag{114}$$
$$=: UB_1 \tag{115}$$

As mentioned, we derive the lower bound for $Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a)$ and finally combine with the upper bound (112) to conclude the upper bound for $\|Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a)\|_\infty$.

$$
\begin{aligned}
&Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) && (116)\\
=&\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle + \gamma \mathbb{E}_{s'}[V^{\pi^*}_{w^*,\tau}(s') - V^{\pi_{t+1}}_{w_{t+1},\tau}(s')] && (117)\\
=&\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle + \gamma \mathbb{E}_{s'}[\langle w^*, V^{\pi^*}_{\tau}(s')\rangle - \langle w_{t+1} - w^* + w^*, V^{\pi_{t+1}}_{\tau}(s')\rangle] && (118)\\
=&\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle + \gamma \mathbb{E}_{s'}[\langle w^*, V^{\pi^*}_{\tau}(s') - V^{\pi_{t+1}}_{\tau}(s')\rangle + \langle w^* - w_{t+1}, V^{\pi_{t+1}}_{\tau}(s')\rangle] && (119)\\
=&\langle w^* - w_{t+1}, \mathbf{r}(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi_{t+1}}_{\tau}(s')]\rangle + \gamma \mathbb{E}_{s'}[V^{\pi^*}_{w^*,\tau}(s') - V^{\pi_{t+1}}_{w^*,\tau}(s')] && (120)\\
=&\langle w^* - w_{t+1}, Q^{\pi_{t+1}}_{\tau}(s,a)\rangle + \gamma \mathbb{E}_{s'}[V^{\pi^*}_{w^*,\tau}(s') - V^{\pi_{t+1}}_{w^*,\tau}(s')] && (121)\\
\geq& - Q_{\tau,max,l_1}\|w^* - w_{t+1}\|_\infty + 0 && (122)\\
\geq& - Q_{\tau,max,l_1}\frac{2}{\tau_w}\| - \mathbf{V}^{\pi^*}_{\tau} - \tau_w \log \kappa_{t+1}\|_\infty \; (\because (77)) && (123)\\
=:& - UB_2 && (124)
\end{aligned}
$$

where $Q_{\tau,max,l_1} = \max_{\pi,s,a}\|Q^{\pi}_{\tau}(s,a)\|_1 = \frac{K(r_{max}+\tau\log|A|)}{1-\gamma}$. In the first inequality, the first term is from Hölder inequality and the second term is from $V^{\pi^*}_{w^*,\tau}(s) \geq V^{\pi_{t+1}}_{w^*,\tau}(s),\ \forall s$ since (π^*, w^*) is a Nash equilibrium in $\mathcal{RG}$. For simplicity, we denote $Q_{\tau,max} = \frac{r_{max}+\tau\log|A|}{1-\gamma}$.

From $-UB_2 \leq Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) \leq UB_1$ (from (112) and (116)),
obtain $\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_{t+1}}_{w_{t+1},\tau}\|_\infty \leq \max\{UB_1, UB_2\}$.

Note that since $r_{max} \leq Q_{\tau,max}$, UB_2 dominates the first term in UB_1. By combining the upper bound (112) and the lower bound (116), conclude the recursive bound for optimality gap of soft Q function as follows.

$$
\begin{aligned}
&\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_{t+1}}_{w_{t+1},\tau}\|_\infty && (125)\\
\leq&\frac{2KQ_{\tau,max}}{\tau_w}\| - \mathbf{V}^{\pi^*}_{\tau} - \tau_w\log\kappa_{t+1}\|_\infty + \gamma\|Q^{\pi^*}_{w^*,\tau} - \tau\log\xi_{t+1}\|_\infty && (126)\\
&+\gamma\max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau\log\xi_{t+1}(s,a)\}\} && (127)
\end{aligned}
$$

Note that the terms in the right-hand side, which are the terms for $t+1$, will be boiled down to the terms for t by using recursive bounds.

Recursive bounds: optimality gap for w For each $k = 1, \dots, K$,

$$
\begin{aligned}
&- V^{\pi^*}_{k,\tau} - \tau_w\log\kappa_{t+1}(k) && (128)\\
=&\beta(-V^{\pi^*}_{k,\tau}) + (1-\beta)(-V^{\pi^*}_{k,\tau}) - \tau_w\beta\log\kappa_t(k) - (1-\beta)(-V^{\pi_t}_{k,\tau}) \; (\because (64)) && (129)\\
=&\beta(-V^{\pi^*}_{k,\tau} - \tau_w\log\kappa_t(k)) + (1-\beta)(-V^{\pi^*}_{k,\tau} + V^{\pi_t}_{k,\tau}). && (130)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&| - V^{\pi^*}_{k,\tau} - \tau_w\log\kappa_{t+1}(k)| && (131)\\
\leq&\beta| - V^{\pi^*}_{k,\tau} - \tau_w\log\kappa_t(k)| + (1-\beta)| - V^{\pi^*}_{k,\tau} + V^{\pi_t}_{k,\tau}| && (132)\\
\leq&\beta| - V^{\pi^*}_{k,\tau} - \tau_w\log\kappa_t(k)| + (1-\beta)\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau} - \tau\log\xi_t\|_\infty && (133)
\end{aligned}
$$

where the last inequality is due to the lemma 15 in Yang et al. [62], applying with $\pi^*(\cdot|s) = \mathrm{softmax}(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau)$ and $\pi_t(\cdot|s) = \mathrm{softmax}(\log\xi_t(s,\cdot))$ and $M = \frac{r_{max}(1+\gamma)+2\tau(1-\gamma)\log|A|}{(1-\gamma)^2}$ following Yang et al. [62].

By taking $\max_k$, we obtain the upper bound for optimality gap of w as follows.

$$
\| - \mathbf{V}^{\pi^*}_{\tau} - \tau_w\log\kappa_{t+1}\|_\infty \leq \beta\| - \mathbf{V}^{\pi^*}_{\tau} - \tau_w\log\kappa_t\|_\infty + (1-\beta)\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau} - \tau\log\xi_t\|_\infty \quad (134)
$$

Recursive bounds: supplementary term

$$Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a) \tag{135}$$
$$= Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau\alpha \log \xi_t(s,a) - (1-\alpha) Q^{\pi_t}_{w_t,\tau}(s,a) \; (\because (62)) \tag{136}$$
$$= Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a) + \alpha (Q^{\pi_t}_{w_t,\tau}(s,a) - \tau \log \xi_t(s,a)) \tag{137}$$

To establish a lower bound for the term $Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a)$,

$$Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a) \tag{138}$$
$$= (Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_{t+1}}_{w_t,\tau}(s,a)) + (Q^{\pi_{t+1}}_{w_t,\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a)) \tag{139}$$
$$\geq Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_{t+1}}_{w_t,\tau}(s,a) \tag{140}$$
$$= \langle w_{t+1} - w_t, Q^{\pi_{t+1}}_{\tau}(s,a) \rangle \tag{141}$$
$$\geq - Q_{\tau,max,l_1} \| w_{t+1} - w_t \|_\infty \; (\because \textit{Hölder} \text{ inequality}) \tag{142}$$
$$\geq - Q_{\tau,max,l_1} \| \log w_{t+1} - \log w_t \|_\infty \; (\because \text{lemma G.1}) \tag{143}$$
$$\geq - 2Q_{\tau,max,l_1} \| \log \kappa_{t+1} - \log \kappa_t \|_\infty \; (\because w_t = \frac{\kappa_t}{\|\kappa_t\|_1} \; \forall t \text{ and 1-Lipschitzness of lse}) \tag{144}$$
$$= - \frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w} \| -\mathbf{V}^{\pi_t}_{\tau} - \tau_w \log \kappa_t \|_\infty \tag{145}$$
$$\geq - \frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w} (\| -\mathbf{V}^{\pi_t}_{\tau} + \mathbf{V}^{\pi^*}_{\tau} \|_\infty + \| -\mathbf{V}^{\pi^*}_{\tau} - \tau_w \log \kappa_t \|_\infty) \tag{146}$$
$$\geq - \frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w} (\frac{M}{\tau} \| Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_t \|_\infty + \| -\mathbf{V}^{\pi^*}_{\tau} - \tau_w \log \kappa_t \|_\infty). \tag{147}$$

The third line is due to performance improvement lemma of NPG with fixed reward (Cen et al. [11], lemma 1), which implies $Q^{\pi_{t+1}}_{w_t,\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a) \geq 0, \; \forall t, s, a$. For completeness, we provide the statement of the lemma below.

Lemma G.2. *(Performance improvement by NPG with fixed scalar reward; adaptation of Lemma 1 in Cen et al. [11])*
For $0 < \eta \leq \frac{1-\gamma}{\tau}$,
$V^{\pi_{t+1}}_{w_{t+1},\tau} - V^{\pi_t}_{w_t,\tau} = \mathbb{E}_{s \sim d^{\pi_{t+1}}_{\mu}}[(\frac{1}{\eta} - \frac{\tau}{1-\gamma}) D_{KL}(\pi_{t+1}(\cdot|s) \| \pi_t(\cdot|s)) + \frac{1}{\eta} D_{KL}(\pi_t(\cdot|s) \| \pi_{t+1}(\cdot|s))]$
where $d^{\pi_{t+1}}_{\mu}$ *is a stationary distribution induced by policy* π_{t+1} *and initial state distribution* μ*,*
i.e., $d^{\pi_{t+1}}_{\mu}(s) = (1-\gamma) \sum_{s_0} \sum_{t=0}^{\infty} \gamma^t Pr^{\pi_{t+1}}(s_t = s | s_0) \mu(s_0)$.
As a result, $Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a) \geq 0$ *for any* s, a.

The last inequality (147) is due to the lemma 15 in Yang et al. [62], applying with $\pi^*(\cdot|s) = softmax(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau)$ and $\pi_t(\cdot|s) = softmax(\log \xi_t(s,\cdot))$.

Plugging (147) into (137) results in the following bound.

$$Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a) \tag{148}$$
$$= Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a) + \alpha (Q^{\pi_t}_{w_t,\tau}(s,a) - \tau \log \xi_t(s,a)) \tag{149}$$
$$\geq - \frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w} (\frac{M}{\tau} \| Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_t \|_\infty + \| -\mathbf{V}^{\pi^*}_{\tau} - \tau_w \log \kappa_t \|_\infty) \tag{150}$$
$$+ \alpha (Q^{\pi_t}_{w_t,\tau}(s,a) - \tau \log \xi_t(s,a)), \tag{151}$$

which implies

$$-(Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \xi_{t+1}(s,a)) \tag{152}$$
$$\leq \frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w} (\frac{M}{\tau} \| Q^{\pi^*}_{w^*,\tau} - \tau \log \xi_t \|_\infty + \| -\mathbf{V}^{\pi^*}_{\tau} - \tau_w \log \kappa_t \|_\infty) \tag{153}$$
$$- \alpha (Q^{\pi_t}_{w_t,\tau}(s,a) - \tau \log \xi_t(s,a)). \tag{154}$$

By taking $\max_{s,a}$ for both sides,

$$\begin{aligned}
&-\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a)-\tau\log\xi_{t+1}(s,a)\} && (155)\\
\leq&\frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w}(\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau}-\tau\log\xi_t\|_\infty+\|-\mathbf{V}^{\pi^*}_\tau-\tau_w\log\kappa_t\|_\infty) && (156)\\
&-\alpha\min_{s,a}\{Q^{\pi_t}_{w_t,\tau}(s,a)-\tau\log\xi_t(s,a)\} && (157)\\
\leq&\frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w}(\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau}-\tau\log\xi_t\|_\infty+\|-\mathbf{V}^{\pi^*}_\tau-\tau_w\log\kappa_t\|_\infty) && (158)\\
&+\alpha\max\{0,-\min_{s,a}\{Q^{\pi_t}_{w_t,\tau}(s,a)-\tau\log\xi_t(s,a)\}\}. && (159)
\end{aligned}$$

Since the last line is non-negative,

$$\begin{aligned}
&\max\{0,-\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a)-\tau\log\xi_{t+1}(s,a)\}\} && (160)\\
\leq&\alpha\max\{0,-\min_{s,a}\{Q^{\pi_t}_{w_t,\tau}(s,a)-\tau\log\xi_t(s,a)\}\} && (161)\\
+&\frac{2Q_{\tau,max,l_1}(1-\beta)}{\tau_w}(\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau}-\tau\log\xi_t\|_\infty+\|-\mathbf{V}^{\pi^*}_\tau-\tau_w\log\kappa_t\|_\infty). && (162)
\end{aligned}$$

G.3 Linear system for optimality gaps

For simplicity, we denote the three optimality gaps and supplementary term as follows.

$$\begin{aligned}
G(\pi_t):=&\|Q^{\pi^*}_{w^*,\tau}-\tau\log\xi_t\|_\infty && (163)\\
G(Q_t):=&\|Q^{\pi^*}_{w^*,\tau}-Q^{\pi_t}_{w_t,\tau}\|_\infty && (164)\\
G(w_t):=&\|-\mathbf{V}^{\pi^*}_\tau-\tau_w\log\kappa_t\|_\infty && (165)\\
H_t:=&\max\{0,-\min_{s,a}(Q^{\pi_t}_{w_t,\tau}-\tau\log\xi_t)\} && (166)
\end{aligned}$$

The recursive bound for these $G(\pi_t), G(Q_t), G(w_t)$ and H_t are summarized below.

$$\begin{cases}
G(Q_{t+1})\leq\frac{2KQ_{\tau,max}}{\tau_w}G(w_{t+1})+\gamma G(\pi_{t+1})+\gamma H_{t+1} & (\because (125))\\
G(\pi_{t+1})\leq\alpha G(\pi_t)+(1-\alpha)G(Q_t) & (\because (83))\\
\qquad\leq(\alpha+(1-\alpha)\gamma)G(\pi_t)+(1-\alpha)(\frac{2KQ_{\tau,max}}{\tau_w}G(w_t)+\gamma H_t) & \\
\qquad(\because \text{by the bound for } G(Q_{t+1}) \text{ above}) & \\
G(w_{t+1})\leq\beta G(w_t)+(1-\beta)\frac{M}{\tau}G(\pi_t) & (\because (134))\\
H_{t+1}\leq\alpha H_t+\frac{2KMQ_{\tau,max}(1-\beta)}{\tau\tau_w}G(\pi_t)+\frac{2KQ_{\tau,max}(1-\beta)}{\tau_w}G(w_t) & (\because (160))
\end{cases}$$

This can be written into the following linear system.

$$\begin{bmatrix}G(\pi_{t+1})\\G(w_{t+1})\\H_{t+1}\end{bmatrix}\leq\begin{bmatrix}\alpha+(1-\alpha)\gamma & \frac{2KQ_{\tau,max}(1-\alpha)}{\tau_w} & (1-\alpha)\gamma\\ \frac{M(1-\beta)}{\tau} & \beta & 0\\ \frac{2KMQ_{\tau,max}(1-\beta)}{\tau\tau_w} & \frac{2KQ_{\tau,max}(1-\beta)}{\tau_w} & \alpha\end{bmatrix}\begin{bmatrix}G(\pi_t)\\G(w_t)\\H_t\end{bmatrix} \tag{167}$$

where $M=\frac{r_{max}(1+\gamma)+2\tau(1-\gamma)\log|A|}{(1-\gamma)^2}, \alpha=1-\frac{\tau\eta}{1-\gamma}, \beta=\frac{1}{\lambda\tau_w+1}$.

Let the transition matrix above be $A(\eta,\lambda)$. Since $A(\eta,\lambda)$ is non-negative matrix, by Perron-Frobenius Theorem [25], $A(\eta,\lambda)$ has an eigenvalue which is equal to spectral radius. The characteristic polynomial is simplified as follows.

$$f(x) = (x-\alpha)\{\underbrace{(x-(\alpha+(1-\alpha)\gamma))(x-\beta)}_{=:f_1(x)} \tag{168}$$

$$-\underbrace{\frac{2KMQ_{\tau,max}}{\tau\tau_w}}_{=:X}(1-\alpha)(1-\beta)\} \tag{169}$$

$$-\underbrace{\gamma\frac{2KMQ_{\tau,max}}{\tau\tau_w}(1-\alpha)(1-\beta)(x+1-2\beta)}_{=:g(x)} \tag{170}$$

Take step size η and λ to satisfy $\alpha = 1-\epsilon$ and $\beta = 1-\epsilon^2$. I.e., take $\eta = \frac{\epsilon(1-\gamma)}{\tau}$ and $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)}$. Note that the polynomial f_1 has zeros at $\alpha+(1-\alpha)\gamma, \beta < 1$. Let $f_2(x) = f_1(x) - X(1-\alpha)(1-\beta)$, then $f_2(1) = (1-\gamma-X)(1-\alpha)(1-\beta)$.

$$X = \frac{2KMQ_{\tau,max}}{\tau\tau_w} < \frac{2K}{\tau\tau_w}\frac{2(r_{max}+\tau\log|A|)}{(1-\gamma)^2}Q_{\tau,max} \tag{171}$$

$$= \frac{4KQ_{\tau,max}^2}{\tau\tau_w(1-\gamma)} \tag{172}$$

$$\leq \frac{1-\gamma}{3} \tag{173}$$

by assumption on τ_w. Therefore, $f_2(1) > 0$ which implies that the zeros of f_2, namely $x_1 \leq x_2$ are less than 1. This implies that the first term of f, i.e. $(x-\alpha)f_2(x)$, has three distinct zeros at $\alpha = 1-\epsilon < \alpha+(1-\alpha)\gamma = 1-\epsilon(1-\gamma) < \beta = 1-\epsilon^2$ for $\epsilon < 1-\gamma$. Furthermore, $(x-\alpha)(f_1(x) - X(1-\alpha)(1-\beta))|_{x=1} = \epsilon f_2(1) = \epsilon^4(1-\gamma-X)$ where X is independent of ϵ. For the last term, $g(1)$ has value $\gamma\frac{4KMQ_{\tau,max}}{\tau\tau_w}\epsilon^5 < \frac{2}{3}\gamma(1-\gamma)\epsilon^5$. For $\epsilon < \frac{1}{\gamma}$, $f(1) = \epsilon^4(1-\gamma-X) - g(1) > \epsilon^4(1-\gamma-\frac{1-\gamma}{3}) - \frac{2}{3}\gamma(1-\gamma)\epsilon^5 > \epsilon^4(\frac{2(1-\gamma)}{3} - \frac{2}{3}\gamma(1-\gamma)\epsilon) > 0$. Note that since $\gamma, \epsilon < 1$, ϵ already satisfies $\epsilon < \frac{1}{\gamma}$. Therefore, x_2, the maximal zero of $(x-\alpha)f_2(x)$ moves to x_2', the maximal zero of $f(x) = (x-\alpha)f_2(x) - g(x)$ which is still $x_2' < 1$(Otherwise, $f(x) \leq 0$). Denote this $x_2' < 1$ as $\rho(\eta,\lambda)$, which is the spectral radius of $A(\eta,\lambda)$.

Specifically, for $x = 1-\frac{\epsilon^2}{2}$, direct calculation results in

$$f(x) = (\frac{1-\gamma}{2} - X)\epsilon^4 - (\frac{2-\gamma}{4} + \frac{3\gamma-1}{2}X)\epsilon^5 + \frac{1}{8}\epsilon^6 \tag{174}$$

$$> (\frac{1-\gamma}{6} - (\frac{2-\gamma}{4} + \frac{3\gamma-1}{2}\frac{1-\gamma}{3})\epsilon)\epsilon^4 \ (\because X < \frac{1-\gamma}{3}) \tag{175}$$

$$= (2(1-\gamma) - (4+5\gamma-6\gamma^2)\epsilon)\frac{\epsilon^4}{12} > 0 \tag{176}$$

by assumption $\epsilon < \frac{2(1-\gamma)}{4+5\gamma-6\gamma^2}$. Therefore, $\rho(\eta,\lambda) < 1-\frac{\epsilon^2}{2} < 1$.

For simplicity, let $\mathbf{x}_t = \begin{bmatrix} G(\pi_t) \\ G(w_t) \\ H_t \end{bmatrix}$. The linear system (167) leads to

$$G(\pi_t), G(w_t), H_t \leq \|\mathbf{x}_t\| \leq \rho(\eta,\lambda)^t\|\mathbf{x}_0\| \tag{177}$$

for some norm. Finally, by (68), (77), (125) and (173), conclude the following.

$$\|\log\pi^* - \log\pi_t\|_\infty \leq \frac{2}{\tau}\|\mathbf{x}_0\|\rho(\eta,\lambda)^t \tag{178}$$

$$\|w^* - w\|_\infty \leq \frac{2}{\tau_w}\|\mathbf{x}_0\|\rho(\eta,\lambda)^t \tag{179}$$

$$\|Q_{w^*,\tau}^{\pi^*} - Q_{w_t,\tau}^{\pi_t}\|_\infty \leq (\frac{\tau(1-\gamma)}{3M} + 2\gamma)\|\mathbf{x}_0\|\rho(\eta,\lambda)^t \tag{180}$$

where $\rho(\eta,\lambda) < 1-\frac{\epsilon^2}{2}$ with choice of $\eta = \frac{\epsilon(1-\gamma)}{\tau}$ and $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)}$.

Remark G.3. *The condition for ϵ is $\epsilon < \epsilon_0 := \min\{\frac{2(1-\gamma)}{4+5\gamma-6\gamma^2}, 1-\gamma\}$ can be simplified with tighter bound. Since $3 \leq 4+5\gamma-6\gamma^2 \leq \frac{121}{24}$ for $\gamma \in (0,1)$, $\frac{2(1-\gamma)}{4+5\gamma-6\gamma^2} \leq \frac{2(1-\gamma)}{3} < 1-\gamma$ and thus, $\epsilon_0 := \min\{\frac{2(1-\gamma)}{4+5\gamma-6\gamma^2}, 1-\gamma\} = \frac{2(1-\gamma)}{4+5\gamma-6\gamma^2}$. In particular, we can take stronger condition $\epsilon < \epsilon_0 := \frac{48(1-\gamma)}{121}$. Condition for λ becomes $\lambda < \frac{\epsilon_0^2}{\tau_w}$, then, from $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)} > \frac{\epsilon^2}{\tau_w}$, λ satisfies the condition for ϵ. Condition for η becomes $\eta < \frac{\epsilon_0(1-\gamma)}{\tau}$, which also satisfies the condition for ϵ by $\epsilon = \frac{\eta\tau}{1-\gamma} < \epsilon_0$.*

H Proof of Theorem 4.3

Now suppose that we can access only an approximate policy evaluation (i.e. estimate for soft Q-function). Let an approximate policy evaluation for given π with reward $\langle w, \mathbf{r}\rangle$ be $\widehat{Q}^{\pi}_{w,\tau}$. We assume that $\|\widehat{Q}^{\pi}_{w,\tau} - Q^{\pi}_{w,\tau}\|_\infty \leq \delta, \;\; \forall \pi, w, s, a$ (Assumption J.2).

We conduct similar analysis in Appendix G with approximate policy evaluation and derive a linear system.

Similar to the appendix G, define auxiliary variables.

$$\widehat{\xi}_0(s,a) := \| \exp(Q^{\pi^*}_{w^*,\tau}(s,\cdot)/\tau)\|_1 \pi_0(a|s) \tag{181}$$

$$\widehat{\xi}_{t+1}(s,a) := (\widehat{\xi}_t(s,a))^{\alpha} e^{\frac{1-\alpha}{\tau}\widehat{Q}^{\pi_t}_{w_t,\tau}(s,a)} \tag{182}$$

$$\widehat{\kappa}_0(k) = \| \exp(-\mathbf{V}^{\pi^*}_{\tau}/\tau_w)\|_1 w_0(k) \tag{183}$$

$$\widehat{\kappa}_{t+1}(k) = (\widehat{\kappa}_t(k))^{\beta} e^{-\frac{1-\beta}{\tau_w}\widehat{\mathbf{V}}^{\pi_t}_{\tau}} \tag{184}$$

$$\pi_t, w_t : \text{policy and weight at } t\text{-th iteration} \tag{185}$$

$$(\pi^*, w^*) : \text{Nash equilibrium in } \mathcal{RG} \tag{186}$$

$$Q^{\pi}_{w,\tau}, V^{\pi}_{w,\tau} : \text{exact policy evaluations} \tag{187}$$

$$\widehat{Q}^{\pi}_{w,\tau}, \widehat{V}^{\pi}_{w,\tau} : \text{approximate policy evaluations} \tag{188}$$

$$\text{I.e., estimators for soft optimal value for given policy } \pi \text{ and reward } \langle w, \mathbf{r}\rangle. \tag{189}$$

$$\widehat{Q}^{\pi}_{\tau}, \widehat{V}^{\pi}_{\tau} : \text{approximate policy evaluations} \tag{190}$$

$$\text{I.e., estimators for soft optimal value for given policy } \pi \text{ and vector reward } \mathbf{r}. \tag{191}$$

$$Q^{*}_{w,\tau} := Q^{\pi^*(w)}_{w,\tau} \triangleq \max_{\pi} Q^{\pi}_{w,\tau} \text{ (soft optimal } Q\text{-value w.r.t. reward } \langle w, \mathbf{r}\rangle) \tag{192}$$

$$V^{*}_{w,\tau} := V^{\pi^*(w)}_{w,\tau} \triangleq \max_{\pi} V^{\pi}_{w,\tau} \text{ (soft optimal } V\text{-value w.r.t. reward } \langle w, \mathbf{r}\rangle) \tag{193}$$

Note that $\pi_t(\cdot|s) = \widehat{\xi}_t(s,\cdot)/\|\widehat{\xi}_t(s,\cdot)\|_1$ also holds for the case with approximate evaluation by the definition of $\widehat{\xi}_t$.

Performance difference of NPG with approximate evaluation
Before deriving the bound for the equation above, we first adapt performance difference lemma with approximate evaluation to our setting.

Performance difference for $V^{\pi_t}_{w_t,\tau}$ with approximate evaluation becomes

$$V^{\pi_{t+1}}_{w_{t+1},\tau}(s) - V^{\pi_t}_{w_t,\tau}(s) \tag{194}$$

$$= V^{\pi_{t+1}}_{w_{t+1},\tau}(s) - V^{\pi_{t+1}}_{w_t,\tau}(s) + V^{\pi_{t+1}}_{w_t,\tau}(s) - V^{\pi_t}_{w_t,\tau}(s) \tag{195}$$

$$\geq \langle w_{t+1} - w_t, \mathbf{V}^{\pi_{t+1}}_{\tau}(s)\rangle - \frac{2}{1-\gamma}\|\widehat{Q}^{\pi_t}_{w_t,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty \;(\because \text{[11], lemma 4}) \tag{196}$$

$$\geq - V_{\tau,max,l_1}\|w_{t+1} - w_t\|_\infty - \frac{2}{1-\gamma}\delta \tag{197}$$

$$= - Q_{\tau,max,l_1}\|w_{t+1} - w_t\|_\infty - \frac{2}{1-\gamma}\delta, \tag{198}$$

and performance difference for $Q^{\pi_t}_{w_t,tau}$ with approximate evaluation becomes

$$Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a) \tag{199}$$

$$=\langle w_{t+1} - w_t, \mathbf{r}(s,a)\rangle + \gamma\mathbb{E}_{s'}[V^{\pi_{t+1}}_{w_{t+1},\tau}(s') - V^{\pi_t}_{w_t,\tau}(s')] \tag{200}$$

$$\geq - Kr_{max}\|w_{t+1} - w_t\|_\infty - \gamma Q_{\tau,max,l_1}\|w_{t+1} - w_t\|_\infty - \frac{2\gamma\delta}{1-\gamma} \tag{201}$$

$$= - Q_{\tau,max,l_1}\|w_{t+1} - w_t\|_\infty - \frac{2\gamma\delta}{1-\gamma}. \tag{202}$$

H.1 Recursive bounds for with approximate evaluation

Recursive bounds with approximate evaluation: optimality gap for policy For each (s,a),

$$Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\log\widehat{\xi}_{t+1}(s,a) \tag{203}$$

$$=Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\alpha\log\widehat{\xi}_t(s,a) - (1-\alpha)\widehat{Q}^{\pi_t}_{w_t,\tau}(s,a)\ (\because (182)) \tag{204}$$

$$=\alpha(Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\log\widehat{\xi}_t(s,a)) + (1-\alpha)(Q^{\pi^*}_{w^*,\tau}(s,a) - \widehat{Q}^{\pi_t}_{w_t,\tau}(s,a)) \tag{205}$$

$$=\alpha(Q^{\pi^*}_{w^*,\tau}(s,a) - \tau\log\widehat{\xi}_t(s,a)) \tag{206}$$

$$+(1-\alpha)(Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_t}_{w_t,\tau}(s,a)) + (1-\alpha)(Q^{\pi_t}_{w_t,\tau}(s,a) - \widehat{Q}^{\pi_t}_{w_t,\tau}(s,a)) \tag{207}$$

Thus,

$$\|Q^{\pi^*}_{w^*,\tau} - \tau\log\widehat{\xi}_{t+1}\|_\infty \leq \alpha\|Q^{\pi^*}_{w^*,\tau} - \tau\log\widehat{\xi}_t\|_\infty + (1-\alpha)\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty + (1-\alpha)\delta. \tag{208}$$

Recursive bounds with approximate evaluation: optimality gap for soft Q function 1) Upper bound
Similar the case with exact evaluation, start from the following identity.

$$Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) \tag{209}$$

$$=\langle w^* - w_{t+1}, \mathbf{r}(s,a)\rangle \tag{210}$$

$$+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\tau\log\|e^{Q^{\pi^*}_{w^*,\tau}(s',\cdot)/\tau}\|_1 - \tau\log\|\widehat{\xi}_{t+1}(s',\cdot)\|_1] \tag{211}$$

$$-\gamma\mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi_{t+1}(\cdot|s')}[Q^{\pi_{t+1}}_{w_{t+1},\tau}(s',a') - \tau\log\widehat{\xi}_{t+1}(s',a')] \tag{212}$$

$$\leq\frac{2Kr_{max}}{\tau_w}\| - \mathbf{V}^{\pi^*}_\tau - \tau_w\log\widehat{\kappa}_{t+1}\|_\infty \tag{213}$$

$$+\gamma\|Q^{\pi^*}_{w^*,\tau} - \tau\log\widehat{\xi}_{t+1}\|_\infty \tag{214}$$

$$+\gamma\max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau\log\widehat{\xi}_{t+1}(s,a)\}\} \tag{215}$$

where the last inequality is derived by the similar logic in the case with exact evaluation.

2) Lower bound
Similar to the case with exact evaluation,

$$Q^{\pi^*}_{w^*,\tau}(s,a) - Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) \geq -\frac{2Q_{\tau,max,l_1}}{\tau_w}\| - \mathbf{V}^{\pi^*}_\tau - \tau_w\log\widehat{\kappa}_{t+1}\|_\infty. \tag{216}$$

Combining the upper bound and the lower bound, obtain the following upper bound for the optimality gap.

$$\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_{t+1}}_{w_{t+1},\tau}\|_\infty \tag{217}$$

$$\leq\frac{2KQ_{\tau,max}}{\tau_w}\| - \mathbf{V}^{\pi^*}_\tau - \tau_w\log\widehat{\kappa}_{t+1}\|_\infty + \gamma\|Q^{\pi^*}_{w^*,\tau} - \tau\log\widehat{\xi}_{t+1}\|_\infty \tag{218}$$

$$+\gamma\max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau\log\widehat{\xi}_{t+1}(s,a)\}\} \tag{219}$$

Recursive bounds with approximate evaluation: optimality gap for w

For each $k = 1, \dots, K$,

$$-V_{k,\tau}^{\pi^*} - \tau_w \log \widehat{\kappa}_{t+1}(k) \tag{220}$$

$$=\beta(-V_{k,\tau}^{\pi^*}) + (1-\beta)(-V_{k,\tau}^{\pi^*}) - \tau_w \beta \log \widehat{\kappa}_t(k) - (1-\beta)(-\widehat{V}_{k,\tau}^{\pi_t}) \; (\because (184)) \tag{221}$$

$$=\beta(-V_{k,\tau}^{\pi^*} - \tau_w \log \widehat{\kappa}_t(k)) + (1-\beta)(-V_{k,\tau}^{\pi^*} + V_{k,\tau}^{\pi_t}) + (1-\beta)(-V_{k,\tau}^{\pi_t} + \widehat{V}_{k,\tau}^{\pi_t}). \tag{222}$$

Then, for the same logic in the case with exact evaluation, obtain the following bound.

$$\| - \mathbf{V}_{\tau}^{\pi^*} - \tau_w \log \widehat{\kappa}_{t+1} \|_\infty \tag{223}$$

$$\leq \beta \| - \mathbf{V}_{\tau}^{\pi^*} - \tau_w \log \widehat{\kappa}_t \|_\infty + (1-\beta) \frac{M}{\tau} \| Q_{w^*,\tau}^{\pi^*} - \tau \log \widehat{\xi}_t \|_\infty + (1-\beta)\delta \tag{224}$$

Recursive bounds with approximate evaluation: supplementary term

$$Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - \tau \log \widehat{\xi}_{t+1}(s,a) \tag{225}$$

$$=Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - \tau\alpha \log \widehat{\xi}_t(s,a) - (1-\alpha)\widehat{Q}_{w_t,\tau}^{\pi_t}(s,a) \; (\because (182)) \tag{226}$$

$$=Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - \tau\alpha \log \widehat{\xi}_t(s,a) - (1-\alpha)(\widehat{Q}_{w_t,\tau}^{\pi_t}(s,a) - Q_{w_t,\tau}^{\pi_t}(s,a) + Q_{w_t,\tau}^{\pi_t}(s,a)) \tag{227}$$

$$=Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - Q_{w_t,\tau}^{\pi_t}(s,a) \tag{228}$$

$$+\alpha(Q_{w_t,\tau}^{\pi_t}(s,a) - \tau\alpha \log \widehat{\xi}_t(s,a)) - (1-\alpha)(\widehat{Q}_{w_t,\tau}^{\pi_t}(s,a) - Q_{w_t,\tau}^{\pi_t}(s,a)) \tag{229}$$

Modify (138) with approximate evaluation as follows.

$$Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - Q_{w_t,\tau}^{\pi_t}(s,a) \tag{230}$$

$$=(Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - Q_{w_t,\tau}^{\pi_{t+1}}(s,a)) + (Q_{w_t,\tau}^{\pi_{t+1}}(s,a) - Q_{w_t,\tau}^{\pi_t}(s,a)) \tag{231}$$

$$\geq Q_{w_{t+1},\tau}^{\pi_{t+1}}(s,a) - Q_{w_t,\tau}^{\pi_{t+1}}(s,a) - Q_{\tau,max,l_1} \| w_{t+1} - w_t \|_\infty - \frac{2\gamma\delta}{1-\gamma} \; (\because (199)) \tag{232}$$

$$=\langle w_{t+1} - w_t, Q_{\tau}^{\pi_{t+1}} \rangle - Q_{\tau,max,l_1} \| w_{t+1} - w_t \|_\infty - \frac{2\gamma\delta}{1-\gamma} \tag{233}$$

$$\geq - Q_{\tau,max,l_1} \| w_{t+1} - w_t \|_\infty - Q_{\tau,max,l_1} \| w_{t+1} - w_t \|_\infty - \frac{2\gamma\delta}{1-\gamma} \; (\because \textit{Hölder} \text{ inequality}) \tag{234}$$

$$\geq - 2Q_{\tau,max,l_1} \| \log w_{t+1} - \log w_t \|_\infty - \frac{2\gamma\delta}{1-\gamma} \; (\because \text{lemma G.1}) \tag{235}$$

$$\geq - 4Q_{\tau,max,l_1} \| \log \widehat{\kappa}_{t+1} - \log \widehat{\kappa}_t \|_\infty - \frac{2\gamma\delta}{1-\gamma} \; (\because w_t = \frac{\widehat{\kappa}_t}{\|\widehat{\kappa}_t\|_1} \; \forall t \text{ and 1-Lipschitzness of lse}) \tag{236}$$

$$= - \frac{4Q_{\tau,max,l_1}(1-\beta)}{\tau_w} \| - \widehat{\mathbf{V}}_{\tau}^{\pi_t} - \tau_w \log \widehat{\kappa}_t \|_\infty - \frac{2\gamma\delta}{1-\gamma} \tag{237}$$

$$\geq - \frac{4Q_{\tau,max,l_1}(1-\beta)}{\tau_w} (\| - \widehat{\mathbf{V}}_{\tau}^{\pi_t} + \mathbf{V}_{\tau}^{\pi_t} \|_\infty + \| - \mathbf{V}_{\tau}^{\pi_t} + \mathbf{V}_{\tau}^{\pi^*} \|_\infty + \| - \mathbf{V}_{\tau}^{\pi^*} - \tau_w \log \widehat{\kappa}_t \|_\infty) \tag{238}$$

$$- \frac{2\gamma\delta}{1-\gamma} \tag{239}$$

$$\geq - \frac{4Q_{\tau,max,l_1}(1-\beta)}{\tau_w} (\delta + \frac{M}{\tau} \| Q_{w^*,\tau}^{\pi^*} - \tau \log \widehat{\xi}_t \|_\infty + \| - \mathbf{V}_{\tau}^{\pi^*} - \tau_w \log \widehat{\kappa}_t \|_\infty) - \frac{2\gamma\delta}{1-\gamma} \tag{240}$$

Plugging (240) into (229), obtain

$$
\begin{aligned}
&Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \widehat{\xi}_{t+1}(s,a) && (241)\\
\geq& -\frac{4Q_{\tau,max,l_1}(1-\beta)}{\tau_w}(\delta + \frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau} - \tau \log \widehat{\xi}_t\|_\infty + \| - \mathbf{V}^{\pi^*}_\tau - \tau_w \log \widehat{\kappa}_t\|_\infty) - \frac{2\gamma\delta}{1-\gamma} && (242)\\
&+\alpha(Q^{\pi_t}_{w_t,\tau}(s,a) - \tau\alpha \log \widehat{\xi}_t(s,a)) - (1-\alpha)\delta && (243)\\
=& -\frac{4Q_{\tau,max,l_1}(1-\beta)}{\tau_w}(\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau} - \tau \log \widehat{\xi}_t\|_\infty + \| - V^{\pi^*}_\tau - \tau_w \log \widehat{\kappa}_t\|_\infty) && (244)\\
&+\alpha(Q^{\pi_t}_{w_t,\tau}(s,a) - \tau\alpha \log \widehat{\xi}_t(s,a)) && (245)\\
&-((1-\alpha)(1+\frac{2\gamma}{\tau\eta}) - \frac{4Q_{\tau,max,l_1}(1-\beta)}{\tau_w})\delta \ (\because \alpha = 1 - \frac{\eta\tau}{1-\gamma}) && (246)
\end{aligned}
$$

By reversing the sign of both sides, taking $\max_{s,a}$ and $\max\{0,\cdot\}$ as in the case with exact evaluation, we obtain the following bound.

$$
\begin{aligned}
&\max\{0, -\min_{s,a}\{Q^{\pi_{t+1}}_{w_{t+1},\tau}(s,a) - \tau \log \widehat{\xi}_{t+1}(s,a)\}\} && (247)\\
\leq& \alpha \max\{0, -\min_{s,a}\{Q^{\pi_t}_{w_t,\tau}(s,a) - \tau \log \widehat{\xi}_t(s,a)\}\} && (248)\\
&+\frac{4KQ_{\tau,max}(1-\beta)}{\tau_w}(\frac{M}{\tau}\|Q^{\pi^*}_{w^*,\tau} - \tau \log \widehat{\xi}_t\|_\infty + \| - \mathbf{V}^{\pi^*}_\tau - \tau_w \log \widehat{\kappa}_t\|_\infty) && (249)\\
&+\max\{0, ((1-\alpha)(1+\frac{2\gamma}{\tau\eta}) - \frac{4KQ_{\tau,max}(1-\beta)}{\tau_w})\}\delta && (250)
\end{aligned}
$$

H.2 Linear system with approximate evaluation

For simplicity, we denote the three optimality gaps and supplementary term as follows, with slight abuse of notation.

$$
\begin{aligned}
G(\pi_t) :=& \|Q^{\pi^*}_{w^*,\tau} - \tau \log \widehat{\xi}_t\|_\infty && (251)\\
G(Q_t) :=& \|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty && (252)\\
G(w_t) :=& \| - \mathbf{V}^{\pi^*}_\tau - \tau_w \log \widehat{\kappa}_t\|_\infty && (253)\\
H_t :=& \max\{0, -\min_{s,a}(Q^{\pi_t}_{w_t,\tau} - \tau \log \widehat{\xi}_t)\} && (254)
\end{aligned}
$$

$$
\begin{cases}
G(Q_{t+1}) \leq \frac{2KQ_{\tau,max}}{\tau_w}G(w_{t+1}) + \gamma G(\pi_{t+1}) + \gamma H_{t+1} & (\because (217))\\
G(\pi_{t+1}) \leq \alpha G(\pi_t) + (1-\alpha)G(Q_t) + (1-\alpha)\delta & (\because (208))\\
\qquad \leq (\alpha + (1-\alpha)\gamma)G(\pi_t) + (1-\alpha)(\frac{2KQ_{\tau,max}}{\tau_w}G(w_t) + \gamma H_t) + (1-\alpha)\delta & \\
\qquad (\because \text{by the bound for } G(Q_{t+1}) \text{ above}) & \\
G(w_{t+1}) \leq \beta G(w_t) + (1-\beta)\frac{M}{\tau}G(\pi_t) + (1-\beta)\delta & (\because (224))\\
H_{t+1} \leq \alpha H_t + \frac{4KMQ_{\tau,max}(1-\beta)}{\tau\tau_w}G(\pi_t) + \frac{4KQ_{\tau,max}(1-\beta)}{\tau_w}G(w_t) & (\because (250))\\
\qquad + \max\{0, ((1-\alpha)(1+\frac{2\gamma}{\tau\eta}) - \frac{4KQ_{\tau,max}(1-\beta)}{\tau_w})\}\delta &
\end{cases}
$$

Therefore, the resulting linear system becomes the following.

$$
\begin{bmatrix} G(\pi_{t+1}) \\ G(w_{t+1}) \\ H_{t+1} \end{bmatrix} \leq \begin{bmatrix} \alpha + (1-\alpha)\gamma & \frac{2KQ_{\tau,max}(1-\alpha)}{\tau_w} & (1-\alpha)\gamma \\ \frac{M(1-\beta)}{\tau} & \beta & 0 \\ \frac{4KMQ_{\tau,max}(1-\beta)}{\tau\tau_w} & \frac{4KQ_{\tau,max}(1-\beta)}{\tau_w} & \alpha \end{bmatrix} \begin{bmatrix} G(\pi_t) \\ G(w_t) \\ H_t \end{bmatrix} + \delta y \qquad (255)
$$

where

$$y = \begin{bmatrix} 1-\alpha \\ 1-\beta \\ \max\{0, (1-\alpha)(1+\frac{2\gamma}{\tau\eta}) - \frac{4KQ_{\tau,max}(1-\beta)}{\tau_w}\} \end{bmatrix}, \tag{256}$$

$$M = \frac{r_{max}(1+\gamma) + 2\tau(1-\gamma)\log|A|}{(1-\gamma)^2}, \alpha = 1 - \frac{\tau\eta}{1-\gamma}, \text{ and } \beta = \frac{1}{\lambda\tau_w + 1}. \tag{257}$$

Let the transition matrix above be $\widehat{A}(\eta, \lambda)$. The characteristic polynomial is simplified as follows.

$$\widehat{f}(x) = (x - (\alpha + (1-\alpha)\gamma))(x-\beta)(x-\alpha) \tag{258}$$

$$- \frac{2KMQ_{\tau,max}(1-\alpha)(1-\beta)}{\tau\tau_w}(x-\alpha) \tag{259}$$

$$- \gamma\frac{4KMQ_{\tau,max}(1-\alpha)(1-\beta)}{\tau\tau_w}(x+1-2\beta) \tag{260}$$

For the same analysis in Appendix G, $\widehat{\rho}(\eta, \lambda)$, the spectral radius of $\widehat{A}(\eta, \lambda)$ satisfies $\widehat{\rho}(\eta, \lambda) < 1 - \frac{\epsilon^2}{2} < 1$ with assumption $X = \frac{2KMQ_{\tau,max}}{\tau\tau_w} \leq \frac{1-\gamma}{3}$ and $\eta = \frac{\epsilon(1-\gamma)}{\tau}$, $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)}$ with $\epsilon < \epsilon_0 := \min\{\frac{2(1-\gamma)}{4+11\gamma-12\gamma^2}, \frac{1}{2\gamma}, 1-\gamma\}$. More tightly, since $3 \leq 4 + 11\gamma - 12\gamma^2 \leq \frac{313}{48}$ for $\gamma \in (0,1)$, $\frac{2(1-\gamma)}{4+11\gamma-12\gamma^2} < 1-\gamma$ and we can take tighter condition $\epsilon \leq \epsilon + 0 := \min\{\frac{96(1-\gamma)}{313}, \frac{1}{2\gamma}\}$. To satisfy the condition for ϵ, required condition for it suffice for η and λ to satisfy the range $\lambda < \frac{\epsilon_0^2}{\tau_w}$ and $\eta < \frac{\epsilon_0(1-\gamma)}{\tau}$.

Using $\widehat{\rho}(\eta, \lambda) < 1 - \frac{\epsilon^2}{2}$ and by letting $\mathbf{x}_t = \begin{bmatrix} G(\pi_t) \\ G(w_t) \\ H_t \end{bmatrix}$ (slightly different from G since this $\mathbf{x}_t$ contains different auxiliary variables), recursive application of the linear system (255) results in the following.

$$\mathbf{x}_t \leq \widehat{A}(\eta, \lambda)^t \mathbf{x}_0 + (I + \cdots + \widehat{A}(\eta, \lambda)^{t-1})\delta y \tag{261}$$

which results in

$$G(\pi_t), G(w_t), H_t \leq \|\mathbf{x}_t\| \leq \widehat{\rho}(\eta, \lambda)^t \|\mathbf{x}_0\| + \frac{1 - \widehat{\rho}(\eta, \lambda)^t}{1 - \widehat{\rho}(\eta, \lambda)}\|y\|\delta \tag{262}$$

$$\leq \widehat{\rho}(\eta, \lambda)^t \|\mathbf{x}_0\| + \frac{1}{1 - \widehat{\rho}(\eta, \lambda)}\|y\|\delta \tag{263}$$

$$\leq \widehat{\rho}(\eta, \lambda)^t \|\mathbf{x}_0\| + \frac{2\|y\|}{\epsilon^2}\delta \tag{264}$$

for some norm. Therefore,

$$\|\log \pi^* - \log \pi_t\|_\infty \leq \frac{2}{\tau}(\|\mathbf{x}_0\|\widehat{\rho}(\eta, \lambda)^t + \frac{2\|y\|}{\epsilon^2}\delta) \tag{265}$$

$$\|w^* - w\|_\infty \leq \frac{2}{\tau_w}(\|\mathbf{x}_0\|\widehat{\rho}(\eta, \lambda)^t + \frac{2\|y\|}{\epsilon^2}\delta) \tag{266}$$

$$\|Q^{\pi^*}_{w^*,\tau} - Q^{\pi_t}_{w_t,\tau}\|_\infty \leq (\frac{\tau(1-\gamma)}{3M} + 2\gamma)(\|\mathbf{x}_0\|\widehat{\rho}(\eta, \lambda)^t + \frac{2\|y\|}{\epsilon^2}\delta) \tag{267}$$

where $\widehat{\rho}(\eta, \lambda) < 1 - \frac{\epsilon^2}{2}$ with choice of $\eta = \frac{\epsilon(1-\gamma)}{\tau}$ and $\lambda = \frac{\epsilon^2}{\tau_w(1-\epsilon^2)}$.

I Proof of Corollaries 4.2 and 4.4

Proof of Corollary 4.2 For each $i = 1, 2, 3$, to achieve $C_i[\rho(\eta, \lambda)]^t \leq \epsilon_{acc}$, it suffices to find t which satisfies $C_i[\rho(\eta, \lambda)]^t < C_i[1 - \frac{\epsilon^2}{2}]^t = \epsilon_{acc}$. Then, $\log C_i + t\log(1 - \frac{\epsilon^2}{2}) = \log \epsilon_{acc}$. Since $\log(1 - \frac{\epsilon^2}{2}) \approx -\frac{\epsilon^2}{2}$ for small ϵ, this results in $t = \frac{2}{\epsilon^2}(\log C_i - \log \epsilon_{acc}) = O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon_{acc}})$.

Proof of Corollary 4.4 For each $i = 1, 2, 3$, to achieve $\widehat{C}_i[\widehat{\rho}(\eta, \lambda)]^t + \widehat{D}_i\delta/\epsilon^2 \leq 2\epsilon_{acc}$ with $\delta \leq \frac{\epsilon^2 \epsilon_{acc}}{\widehat{D}_i}$, it suffices to find t which satisfies $\widehat{C}_i[\widehat{\rho}(\eta, \lambda)]^t + \widehat{D}_i\delta/\epsilon^2 < \widehat{C}_i[1 - \frac{\epsilon^2}{2}]^t + \epsilon_{acc} = 2\epsilon_{acc}$, i.e., $\widehat{C}_i[1 - \frac{\epsilon^2}{2}]^t = \epsilon_{acc}$. For the same logic in the case of exact policy evaluation, iteration complexity becomes at most $O(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon_{acc}})$.

The derivation and analysis for sample complexity is in Appendix J.

J Sample Complexity under Approximate Policy Evaluation

Assumption J.1. *We assume access to a generative model that allows sampling next states from $P(\cdot \mid s, a)$ for any (s, a), as in Li et al. [28].*

Assumption J.2. *For any given error bound δ, policy π and weight $w \in \Delta^K$, we have access to an estimate $\widehat{Q}^\pi_{w,\tau}$ and $\widehat{Q}^\pi_k$ of value functions $Q^\pi_{w,\tau}$ and Q^π_k, which satisfy $\|\widehat{Q}^\pi_{w,\tau} - Q^\pi_{w,\tau}\|_\infty < \delta$ and $\|\widehat{Q}^\pi_k - Q^\pi_k\|_\infty < \delta$, respectively.*

As discussed in Li et al. [28], Assumption J.2 can be achieved for any fixed π and any k-th objective within error $\|\widehat{Q}^\pi_{k,\tau} - Q^\pi_{k,\tau}\|_\infty \leq \delta$ with high probability if the number of samples per each state-actions pair exceeds the order of $\tilde{O}\left(\frac{1}{(1-\gamma)^3\delta^2}\right)$, under Assumption J.1. By employing fresh samples for the policy evaluation for each objective at every iteration and using union bound, $\|\widehat{Q}^\pi_\tau - Q^\pi_\tau\|_\infty = \max_k \|\widehat{Q}^\pi_{k,\tau} - Q^\pi_{k,\tau}\|_\infty \leq \delta$ is achieved with high probability if the number of samples per each state-actions pair is exceeds the order of $\tilde{O}\left(\frac{K}{(1-\gamma)^3\delta^2}\right)$. In particular, since $Q^\pi_{w,\tau}$ is a soft value function with a scalar reward $\langle w, \mathbf{r}\rangle$, we need samples exceeding the order of $\tilde{O}\left(\frac{1}{(1-\gamma)^3\delta^2}\right)$ to estimate. Since $\tilde{O}\left(\frac{K}{(1-\gamma)^3\delta^2}\right)(\because\ w$ update requires vector value $V^{\pi_t}_\tau)$ + $\tilde{O}\left(\frac{1}{(1-\gamma)^3\delta^2}\right)(\because\ \pi$ update requires scalar value $Q^{\pi_t}_{w_t,\tau})$ = $\tilde{O}\left(\frac{K}{(1-\gamma)^3\delta^2}\right)$, we used $\tilde{O}\left(\frac{K}{(1-\gamma)^3\delta^2}\right)$ in Corollary 4.4.

K Discussion on Two-player Zero-sum Markov Games

We provide a detailed discussion on how our framework relates to the literature on two-player zero-sum Markov games (2p0s MGs).

Problem Settings. Our target problem, entropy-regularized max–min MORL, can be viewed as a two-player zero-sum Markov game in which the minimization player (the adversary) has a state-independent policy w, and the transition is determined solely by the maximization player (the learner), since $P(s'|s, a)$ is independent of w. We also aim to find an ϵ_{acc}-QRE, as considered in several prior works.

Algorithmic Settings. Table 4 summarizes the algorithmic settings of our method and existing studies on two-player zero-sum Markov games. Here, "symm" denotes symmetric regularization and learning-rate structures, while "asym" denotes asymmetric ones. In 2p0s MGs, symmetric regularization is natural because both players use policies as their strategies with cumulative entropy $\tilde{H}(\pi)$ on the order of $\frac{\log|\mathcal{A}|}{1-\gamma}$. In contrast, in ERAM, the adversary uses a weight w as its strategy with non-cumulative entropy $\tilde{H}(w)$ on the order of $\log K$, which justifies the use of asymmetric regularization coefficients.

Techniques for Proof. We carefully reviewed the proof of Theorem 1 in Cen et al. [12] and found it to be applicable to the MORL setting under symmetric learning rates and regularization ($\eta = \lambda$, $\tau = \tau_w$) by using $\frac{1}{1-\gamma}\tilde{H}(w)$ or by absorbing $\frac{1}{1-\gamma}$ into τ_w, thereby yielding asymmetric coefficients. Nevertheless, a direct application remains challenging because our setting lacks these symmetric conditions. Our proof instead relies on asymmetric learning rates for last-iterate convergence (see Section 4) and asymmetric regularization (see Algorithmic Settings above).

Table 4: Comparison of algorithmic settings in two-player zero-sum Markov games.

Reference	Entropy Reg.	Reg. Coef.	Base Method	Policy Param.	Learning Rates
Daskalakis et al. [15]	×	none	PG	direct	asym
Zeng et al. [64]	✓	symm	PG	softmax	asym
Wei et al. [58]	✓	none	OGDA	direct	symm
Cen et al. [12]	✓	symm	OMWU	direct	symm
ERAM (Ours)	✓	asym	NPG, MD	softmax	asym

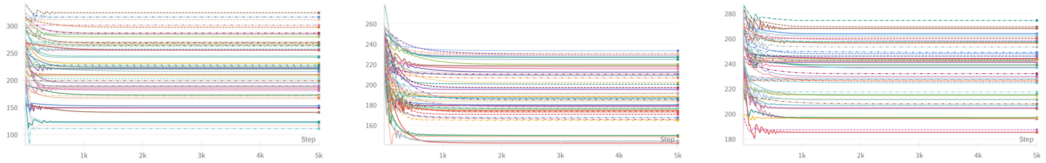
The result of Cen et al. (2023) applies to MORL in the symmetric case because its key step, Eq. (14), has a state-dependent right-hand side (RHS) that is ultimately bounded by $\|\cdot\|_{\Gamma_{(p)}}$, removing all state dependence. Hence, both the state-dependent policy in MGs and the state-independent w in MORL satisfy the same recursive bounds as in Lemma 2–4 of Cen et al. (2023). In our setting, however, since w is already state-independent and there is only one RL learner, we find that adapting the proof in Cen et al. (2022) is sufficient for our analysis.

Convergence Speed. Cen et al. (2023) provides a convergence error bound of $O((1-(1-\gamma)\eta\gamma/4)^t)$, while our method achieves at most $O((1-\epsilon^2/2)^t) \leq \mathcal{O}((1-\lambda\tau_w/4)^t)$. The proof in Appendix G shows that if we take $\epsilon \geq (1-\gamma)^2$, we have $1-\lambda\tau_w/4 \leq 1-(1-\gamma)\eta\gamma/4$. Hence, our convergence rate is comparable to that of prior work and does not conflict with their results. (Note that $\epsilon \geq (1-\gamma)^2$ does not violate $\epsilon \in (0, \epsilon_0 = 48(1-\gamma)/121)$ (Remark G.3), for γ sufficiently close to 1.) This discussion highlights that while the mathematical foundation of ERAM aligns closely with existing analyses in two-player zero-sum Markov games, our asymmetric learning-rate structure and the non-monotone nature of value gradients in RL motivate a distinct line of analysis tailored to the MORL framework.

L Supplementary Material for Numerical Results in Tabular MOMDPs

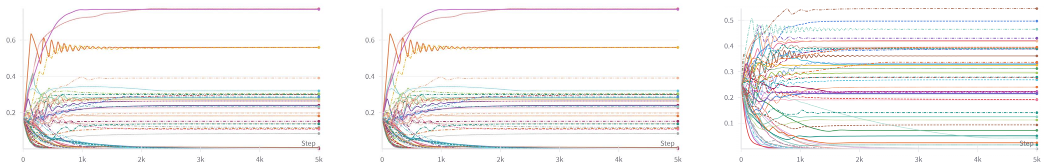
For each tabular MOMDP, transition matrix and reward function is randomly generated. The reward function has range $[1, 20]$. For all experiment, we used $\gamma = 0.95$, $\tau = \tau_w = 0.05$, $\eta = 0.01$ and $\lambda = 0.0001$. The following graphs show last-iterate convergence behaviors for value $V^{\pi_t}_{w_t,\tau}$ and weight w_t. For simplicity, we report $w_t(1)$, the first element of the weight w_t at iteration t.

The figures below show the last-iterate convergence behavior of values ($V^{\pi_t}_{w_t,\tau}$) and the first component of weights ($w_t(1)$) resulting from Algorithm 2.



$(|S|, |A|, K) = (2, 2, 2)$ $(|S|, |A|, K) = (3, 3, 6)$ $(|S|, |A|, K) = (4, 4, 4)$

Figure 5: Last-iterate convergence of values in three types of MOMDPs, each with 50 randomly generated instances.



$(|S|, |A|, K) = (2, 2, 2)$ $(|S|, |A|, K) = (3, 3, 6)$ $(|S|, |A|, K) = (4, 4, 4)$

Figure 6: Last-iterate convergence of $w(1)$ in three types of MOMDPs, each with 50 randomly generated instances.

We also evaluated ARAM in tabular settings. To demonstrate the effectiveness of ARAM, we considered two types of tabular MOMDPs with an increased number of objectives ($K = 10$): $(|S|, |A|, K) = (2, 2, 10)$ and $(4, 4, 10)$. We used $\gamma = 0.95$, $\tau = \tau_w = 0.05$, $\eta = 0.01$ and $\lambda = 0.0001$. Similar to ERAM, the Nash gap of ARAM decreases rapidly over time in both environments, as shown in Figure 7. Each curve represents the average over 50 randomly generated instances, with shaded areas showing standard deviation.

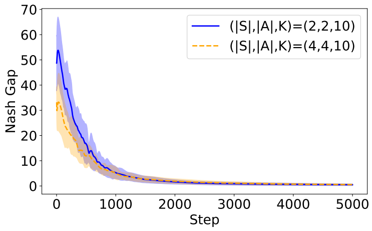


Figure 7: Nash gap for ARAM

M More on the Traffic Signal Control Experiment

M.1 Traffic signal control environment

We evaluate our method in a traffic signal control simulation environment with three scenarios: Base-4, Asym-4, and Asym-16. Each scenario simulates a four-way intersection with four roads (North, East, South, and West), where each road consists of four lanes. Vehicles arrive from the four directions with specified inflow proportions and can proceed straight, turn left, or turn right.

In the Base-4 scenario, 75% of arriving vehicles proceed straight, and the remaining 25% make left or right turns with equal probability. Among the straight-going vehicles, the proportions from West to East, East to West, North to South, and South to North are $[0.1, 0.1, 0.4, 0.4]$. The reward is 4-dimensional, where each element represents the negative total waiting time on a road. The simulation includes 10,000 vehicles and is trained for 100,000 time steps.

The Asym-4 scenario shares the same reward structure as Base-4 but introduces asymmetry in the traffic inflow, which is set to $[0.4, 0.1, 0.4, 0.1]$ across the four directions. Additionally, the turning probabilities vary by incoming direction to better reflect realistic traffic patterns. The scenario uses 4,000 vehicles and is also trained for 100,000 time steps.

The Asym-16 scenario uses the same asymmetric inflow as Asym-4 but increases the granularity of the reward by assigning one reward per lane, resulting in a 16-dimensional reward vector. Each element corresponds to the negative waiting time of a specific lane. This scenario includes 4,000 vehicles and is trained for 200,000 time steps.

For completeness, we provide the max-min performance under the metric of Park et al. [37] in the following table (see Table 5). Let π^* denote the max-min optimal policy, and let π^{seed_i} denote the final policy obtained from training with a fixed seed $seed_i$. The metric used in the prior work [37] averages over all seeds and then takes the minimum, i.e., $\min_k \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\pi^{seed_i}}[\sum_t \gamma^t r_k(s_t, a_t)]$. In contrast, our metric computes the max-min performance for each seed and then averages the results, i.e., $\frac{1}{n}\sum_{i=1}^{n} \min_k \mathbb{E}_{\pi^{seed_i}}[\sum_t \gamma^t r_k(s_t, a_t)]$. If $\pi^{seed_1} = \cdots = \pi^{seed_n} = \pi^*$, the value vectors $\mathbf{V}^{\pi^{seed_i}}$ are identical to $\mathbf{V}^{\pi^*}$, and both metrics yield the same measurement. However, if $\pi^{seed_1}, \cdots, \pi^{seed_n}$ are not exactly the same, the metric in the prior work averages returns that are from different policies, which does not accurately reflect each learned policy's true performance. Thus, we evaluate the max-min performance of each policy separately and then average the results.

Environments	ARAM	ERAM	Park et al. [37]	GGF-PPO	GGF-DQN	Avg-DQN
Base-4	**-1140**	-1387	-1455	-1603	-1838	-2774
Asym-4	-2589	**-2568**	-3510	-3094	-2670	-4245
Asym-16	**-14399**	-15259	-17754	-19569	-16477	-27499

Table 5: Max-min performance in traffic signal control. Bold: best; underline: second-best.

M.2 Experimental Setup

The PPO hyperparameters are listed in the table below. We use the default network architecture and optimizer settings provided by Stable-Baselines3 [43]. For entries with multiple values, the best-performing one was selected based on validation performance.

PPO hyperparameter	value
entropy coefficient	1e-6
value loss coefficient	0.5
gae coefficient	0.95
clip range	0.2
optimizer	Adam
hidden layer sizes for actor network	[64, 64]
hidden layer sizes for critic network	[64, 64]
activation function	Tanh
epochs	2, 4, 6, 8
rollout steps	64, 128
batch size	16, 32
learning rate	0.001, 0.002, 0.003

Table 6: PPO hyperparameters for ERAM and ARAM

The main hyperparameters for ERAM and ARAM are λ and $\beta = \frac{1}{\lambda\tau_w+1}$ (i.e. τ_w), which appear in the closed-form updates of w in both ERAM (13) and ARAM (40). The best-performing hyperparameters were selected for each traffic scenario from the search over $\beta \in \{0.01, 0.25, 0.33, 0.5, 0.67, 0.75, 0.99\}$ and $\lambda \in \{0.001, 0.002, 0.003, 0.01, 0.02, 0.03, 0.1, 0.2, 0.3\}$, and their influence is analyzed in the ablation study. In addition, all experiments were conducted on a machine equipped with two Intel Xeon Gold 6238R CPUs.

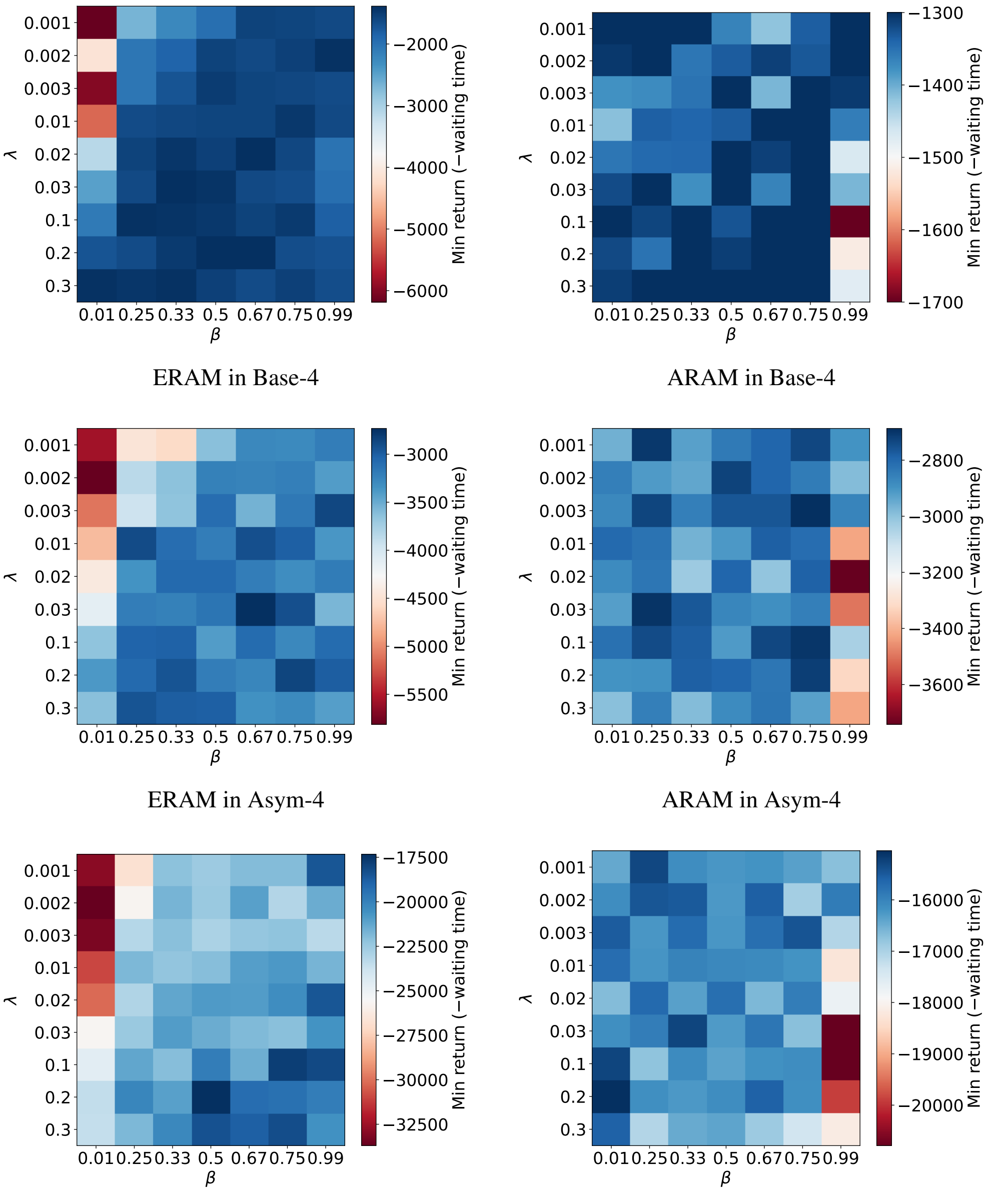
In each traffic scenario, the selected hyperparameters are listed in the following order:
(epochs, rollout steps, batch size, learning rate, λ, β).

	ERAM	ARAM
Base-4	(8, 128, 32, 0.001, 0.2, 0.67)	(8, 128, 32, 0.001, 0.1, 0.67)
Asym-4	(4, 128, 32, 0.003, 0.03, 0.67)	(4, 128, 32, 0.003, 0.03, 0.25)
Asym-16	(8, 128, 32, 0.001, 0.2, 0.5)	(8, 128, 32, 0.001, 0.2, 0.01)

Table 7: Selected hyperparameters for ERAM and ARAM in each traffic scenario

M.3 Ablation study

As mentioned above, we conducted an ablation study on λ and $\beta = \frac{1}{\lambda\tau_w+1}$ (then, $\tau_w = (\frac{1}{\beta} - 1)/\lambda$ is automatically determined), which are the main components of the closed-form update rules for w in both ERAM (13) and ARAM (40). Figure 8 presents the ablation study on λ and β for ERAM and ARAM across different traffic scenarios. Each heatmap visualizes the minimum return obtained for varying (λ, β) combinations. In each heatmap, red indicates lower minimum return, while blue indicates higher minimum return.



ERAM in Base-4

ARAM in Base-4

ERAM in Asym-4

ARAM in Asym-4

ERAM in Asym-16

ARAM in Asym-16

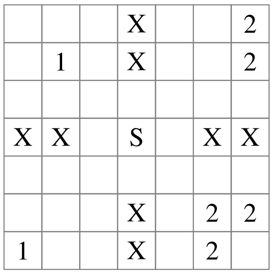
Figure 8: Ablation results on λ and β for ERAM and ARAM across traffic scenarios

N Additional Experimental Results

	ARAM	ERAM	Park et al. [37]	GGF-PPO	GGF-DQN	Avg-DQN
Spec. Cons.	31	27	27	27	22	4
MO-Reacher	25.27	25.13	23.54	24.32	23.90	22.44
Four Room	1.80	1.56	1.02	1.47	0.02	0.12

Table 8: Max-min performance in species conservation environment, MO-Reacher environment, and Four-room environment

The Species Conservation environment (SC) [51], a commonly used benchmark in MORL, aims to promote the conservation of species in an ecological simulation via multi-objective reinforcement learning. This environment includes two species: sea otters (an endangered predator) and northern abalone (their prey). The state space captures population information, and the action space consists of five discrete actions. The agent aims to achieve fair conservation of both species by treating their population levels as a vector-valued reward.



			X			2
	1		X			2
X	X		S		X	X
			X		2	2
1			X		2	

Figure 9: The map for Four-Room environment

MO-Reacher environment (MR) [20] is a multi-objective extension of Reacher [56]. This environment has a 6-dimensional observation space containing the sine and cosine values of the central and elbow joint angles, as well as their angular velocities. It has a discrete action space consisting of torques applied to the central and elbow joints, where each torque can take one of three values: -1, 0, or 1. As an extension of the standard Reacher environment, MO-Reacher includes four targets. The reward function is defined based on the distance between the tip of the arm and each target as follows: $r_i = 1 - 4\|(\text{tip's position}) - (\text{target } i\text{'s position})\|^2,\ i = 1, 2, 3, 4$.

The Four-Room (FR) environment [20] contains two types of collectible items, labeled 1 and 2. Figure 9 shows the map used in the Four-Room environment. The agent starts at the center of the map, marked as "S". Cells marked "X" represent walls. Each episode terminates after a maximum of 200 steps. The agent collects items 1 and 2 throughout the episode, and the numbers of each collected item constitute the two-dimensional objective reward.

In all environments, we trained for a total of 100,000 timesteps. Table 8 demonstrates the superior max-min performance of our algorithms in all environments.

O Limitations and Broader Impacts

The max-min criterion works best when objectives are homogeneous, such as having the same units. Extending it to heterogeneous objectives is an interesting direction for future work. While our theory focuses on softmax policy parameterization, analyzing last-iterate convergence under more general settings, such as linear function approximation, remains a valuable research direction.

Our work on max-min MORL may have several broader impacts. The proposed algorithm can be applied to real-world resource allocation problems where fairness or robustness across competing objectives is critical, such as transportation, healthcare, and public infrastructure. Moreover, the max-min criterion may contribute to more robust preference modeling in RL fine-tuning of large language models (LLMs), especially in the context of preference alignment. In such applications, max-min training can mitigate the influence of outlier preferences and help ensure consistency across diverse feedback signals. Because our algorithms are both memory-efficient and computation-efficient, it may be particularly suitable for training large-scale models under practical resource constraints. While our method is general and does not directly involve deployment, we note that any reinforcement learning system deployed in sensitive domains should be carefully audited for fairness, safety, and long-term behavior. We believe our approach supports these goals by improving robustness in multi-objective decision-making.