Safe and Balanced: A Framework for Constrained Multi-Objective Reinforcement Learning

Shangding Gu, *Graduate Student Member, IEEE*, Bilgehan Sel, Yuhao Ding, Lu Wang, Qingwei Lin, *Member, IEEE*, Alois Knoll, *Fellow, IEEE*, and Ming Jin

***Abstract*—In numerous reinforcement learning (RL) problems involving safety-critical systems, a key challenge lies in balancing multiple objectives while simultaneously meeting all stringent safety constraints. To tackle this issue, we propose a primal-based framework that orchestrates policy optimization between multi-objective learning and constraint adherence. Our method employs a novel natural policy gradient manipulation method to optimize multiple RL objectives and overcome conflicting gradients between different objectives, since the simple weighted average gradient direction may not be beneficial for specific objectives due to misaligned gradients of different objectives. When there is a violation of a hard constraint, our algorithm steps in to rectify the policy to minimize this violation. Particularly, We establish theoretical convergence and constraint violation guarantees, and our proposed method also outperforms prior state-of-the-art methods on challenging safe multi-objective RL tasks.**

***Index Terms*—Constrained reinforcement learning, multi-objective reinforcement learning, gradient manipulation.**

I. Introduction

REINFORCEMENT Learning (RL) has made significant strides and is used widely in various domains [1], e.g., robotics [2], [3], autonomous driving [4], [5], large language model [6], and finance [7]. However, a significant challenge arises when a policy must address multiple objectives within a single task or manage multiple tasks concurrently. Direct optimization of scalarized objectives can lead to suboptimal performance, with the optimizer often struggling to make progress, resulting in a considerable decline in learning performance [8]. A significant cause of this issue is the phenomenon of conflicting gradients [9]. Here, gradients associated with different objectives may vary in scale, potentially leading the largest gradient to dominate the update. Moreover, they might point in different directions, i.e., $\nabla f_i(\pi)^\top \nabla f_j(\pi) < 0, i \neq j, i, j \in [m] = \{1, \ldots, m\}$, causing the performance of one objective to deteriorate during the optimization of another. While recent studies have shown that linear scalarization can be competitive [10], it may fall short when faced with safety-critical constraints. Indeed, ensuring the safe application of RL algorithms in real-world settings, especially those dealing with multiple objectives, is paramount [11]. This study seeks to answer the key question:

How can we balance each objective while ensuring safety constraints?

Addressing this problem, akin to a multi-dimensional tug of war, requires nuance. Each objective is a team pulling in its own direction, yet confined by the boundaries of safety—a balancing act of objectives and safety. Inspired by this dynamic, we devise a comprehensive framework for Constrained Multi-Objective RL (CMORL) using gradient manipulation and constraint rectification. It operates in three stages: (1) Estimating Q-functions from the existing policy. (2) If all constraints are satisfactorily met, the policy is updated via the manipulated natural policy gradient (NPG) of multiple objectives to minimize the gradient conflicts. (3) If not, the policy is updated following the NPG of the unsatisfied constraint. These steps are iteratively repeated until convergence is achieved.

In this framework, we provide a theoretical analysis, including convergence analysis and violation guarantee analysis. Using the insights from this analysis, we develop a practical algorithm to manage multi-objective RL while ensuring safety during learning. We further deploy our algorithm on safe multi-objective tasks in the MuJoCo environment [12] and compare our method with the state-of-the-art (SOTA) safe baseline, CRPO [13], and SOTA safe multi-objective RL methods, such as LP3 [11]. Our experimental results suggest that our method outperforms CRPO and LP3 in striking a balance between reward performance and safety violation.

Our study offers several significant contributions to the field of safe multi-objective RL, which are delineated as follows: (1) A novel framework for safe multi-objective RL, wherein a comprehensive analysis of both theoretical convergence and constraint

Received 5 June 2024; revised 7 December 2024; accepted 31 December 2024. Date of publication 14 January 2025; date of current version 3 April 2025. This work was supported in part by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant 945539 (Human Brain Project SGA3), in part by NSF under Grant ECCS-233177 and Grant IIS-2312794, in part by the Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, and in part by the Commonwealth Cyber Initiative. Recommended for acceptance by W. Zhang. *(Shangding Gu, Bilgehan Sel, and Yuhao Ding contributed equally to this work.) (Corresponding authors: Ming Jin; Shangding Gu.)*

Shangding Gu is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA, and also with the Department of Informatics, Technical University of Munich, 85748 Munich, Germany (e-mail: shangding.gu@berkeley.edu).

Bilgehan Sel and Ming Jin are with the Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: bsel@vt.edu; jinming@vt.edu).

Yuhao Ding is with the Cubist Systematic Strategies, New York City, NY 10036 USA (e-mail: yuhao.ding3@gmail.com).

Lu Wang and Qingwei Lin are with the Microsoft Research Asia, Beijing 100080, China (e-mail: wlu@microsoft.com; qlin@microsoft.com).

Alois Knoll is with the Department of Informatics, Technical University of Munich, 85748 Munich, Germany (e-mail: knoll@mytum.de).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3528944, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3528944

violation guarantees is conducted. (2) The development of a benchmark grounded in MuJoCo environments (Named *Safe Multi-Objective MuJoCo*), aimed at scrutinizing the efficacy of safe multi-objective learning. (3) The superior performance by our proposed method in terms of striking a balance between safety concerns and the accomplishment of multiple reward objectives, as evidenced across numerous challenging tasks within the realm of safe multi-objective RL.

Novelty Discussion Our framework is developed based on CRPO [13] and NPG [14]. Regarding its novelty, we have added a comparative discussion of our method with CRPO [13] and NPG [14].

Main Challenges: The problem we address involves balancing multi-objective optimization while ensuring learning safety. There are two major challenges: 1. Balancing multi-objective optimization while ensuring safety. 2. Providing theoretical guarantees for safe multi-objective optimization. However, CRPO and NPG do not consider these settings. To address these challenges, we propose the first primal-based safe multi-objective optimization framework and conduct comprehensive experiments to evaluate the effectiveness of our algorithm.

Algorithm Novelty: Neither CRPO nor NPG consider safe multi-objective optimization, despite the importance and urgency of these problems when deploying RL in real-world applications. It is critical to balance each objective during policy learning, which CRPO and NPG also do not address. We have designed a novel algorithm to handle a safe multi-objective optimization problem and balance multiple objectives and safety. *Theoretical Analysis Novelty:* We provide rigorous proofs on how to guarantee safety and balance multiple objective optimizations in our theoretical analysis. CRPO and NPG do not offer such guarantees. *Experiment Contributions:* First, we designed a safe multi-objective benchmark and conducted experiments to evaluate the effectiveness of our algorithm. CRPO and NPG do not provide such benchmarks. Second, we compared our algorithm with the SOTA safe multi-objective RL algorithm, LP3. The experimental results demonstrate that our algorithm performs better than the SOTA baselines.

II. Related Work

In recent years, numerous methods are proposed to help deploy RL in real-world applications [1], [11], [15], [16], which try to solve the safe exploration problem [15] or satisfy multi-objective requirements during RL exploration [17] from the perspective of safe or multi-objective RL.

Safe Reinforcement Learning: Safe RL has gained significant attention as it helps address learning safety problems during RL deployment in real-world applications. Safe RL can be considered as a constrained optimization problem [15]. For example, several safe RL methods leverage Gaussian Processes to model the safe state space during exploration [18], [19], [20], [21]. In contrast to modeling the safe state, some safe RL methods attempt to search for a safe policy within the constrained action space [22], [23], [24], [25], [26], [27], e.g., based on formal methods, the exploration action is verified via temporal logic verification during exploration [25]. Furthermore, by optimizing the average cumulative cost of each trajectory, several constrained policy optimization-based methods are proposed, such as CPO [28], PCPO [29], RCPO [30], PDPG [31] and CRPO [13].

Multi-Objective Reinforcement Learning (MORL): There are two settings in MORL [11]. The first involves a single policy in multi-objective optimization, while the second involves a multi-policy set that satisfies multi-objective requirements. Most MORL methods are developed based on the first setting, where a single policy needs to meet multiple objective conditions simultaneously [32]. Additionally, various multi-objective learning methods are proposed to optimize policy performance, such as multi-objective learning as a bargaining game [33], Cagrad [34], the Multiple-Gradient Descent Algorithm (MGDA) [35], and PCGrad [9]. Methods in the second setting attempt to learn a complete set of Pareto frontiers and use a posterior selection to satisfy multi-objective requirements [36]. Examples include MORL optimization based on manifold space to find better solutions on the Pareto frontier [37], [38].

The methods mentioned above address RL safety or multi-objective requirements separately without considering both aspects simultaneously. Our focus is on achieving safe MORL, which involves ensuring exploration safety in multi-objective RL settings. The most similar work to ours is the Learning Preferences and Policies in Parallel (LP3) algorithm [11], which is proposed based on the Multi-Objective Maximum Posterior Policy optimization (MO-MPO) [32]. In this approach, a supervised learning algorithm is used to learn preferences, and then a policy is trained based on Lagrangian optimization. However, their method heavily depends on Q-estimation, which may not accurately represent safe preferences; the gradient conflict between each objective is not analyzed, and neither convergence analysis nor safety violation guarantees are provided. In contrast to LP3 [11], we propose a primal-based framework that can balance policy optimization between multi-objective learning and constraint satisfaction based on conflict-averse NPG. In our approach, the conflict gradient is analyzed between each objective performance, and convergence analysis and safety violation guarantees are provided based on gradient manipulation and constraint rectification.

III. Preliminaries and Problem Formulation

A. Multi-Objective RL (MORL)

A MORL is a tuple $(\mathcal{S}, \mathcal{A}, \{r_i\}_{i=1}^{m}, \mathrm{P}, \rho, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces; $r_i : \mathcal{S} \times \mathcal{A} \to [0, r_{\max}]$ is the reward function; $m \geq 2$ denotes the number of objectives; $\mathrm{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel, with $\mathrm{P}(s' \mid s, a)$ denoting the probability of transitioning to state s' from previous state s given action a; $\rho : \mathcal{S} \to [0, 1]$ is the initial state distribution; and $\gamma \in (0, 1)$ is the discount factor. A policy $\pi \in \Pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is a mapping from the state space to the space of probability distributions over the actions, with $\pi(\cdot \mid s)$ denoting the probability of selecting action a in state s. When the associated Markov chain $\mathrm{P}(s' \mid s) = \sum_{\mathcal{A}} P(s' \mid s, a)\pi(a \mid s)$ is ergodic, we denote μ_π as the stationary distribution of this MDP, i.e. $\int_{\mathcal{S}} \mathrm{P}(s' \mid s)\mu_\pi(ds) = \mu_\pi(s')$. Moreover, we define the visitation measure induced by the policy π as $\nu_\pi(s, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{P}(s_t = s, a_t = a)$.

For a given policy π and a reward function r_i, we define the state value function as $V_i^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \mid s_0 = s, \pi]$, the state-action value function as $Q_i^\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \mid s_0 = s, a_0 = a, \pi]$, the advantage function as $A_i^\pi(s,a) = Q_i^\pi(s,a) - V_i^\pi(s)$, and the expected total reward function $f_i(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)] = \mathbb{E}_\rho[V_i^\pi(s)] = \mathbb{E}_{\rho \cdot \pi}[Q_i^\pi(s,a)]$.

In MORL, we aim to find a single optimal policy that maximizes multiple expected total reward functions simultaneously, termed as

$$\max_{\pi \in \Pi} \boldsymbol{F}(\pi) = (f_1(\pi), \ldots, f_m(\pi))^\top . \quad (1)$$

B. Constrained Multi-Objective RL (CMORL)

The CMORL problem refers to a formulation of MORL that involves additional *hard constraints* that restrict the allowable policies. The constraints take the form of costs that the agent may incur when taking actions at certain states, denoted by the functions $r_{m+1}, \ldots, r_{m+p}$. Each of these cost functions maps a tuple (s,a) to a corresponding cost value. The function $f_{m+i}(\pi)$ represents the expected total cost incurred by the agent with respect to cost function r_{m+i}. The objective of the agent in CMORL is to solve a multi-objective RL problem subject to the aforementioned hard constraints:

$$\max_{\pi \in \Pi} \boldsymbol{F}(\pi), \text{ s.t. } f_i(\pi) \le c_i, \forall i = m+1, \ldots, m+p, \quad (2)$$

where c_i is a fixed limit for the i-th constraint. Here, we overload the max operator to imply pareto optimal policies to handle the vector objective function. We define the safety set $\Pi_{\text{safe}} = \{\pi \in \Pi \mid f_i(\pi) \le c_i, \forall i = m+1, \ldots, m+p\}$, and the optimal policy $\pi^* = \arg\max_{\pi \in \Pi_{\text{safe}}} \boldsymbol{F}(\pi)$ for CMORL in (2). In practice, a convenient way to solve RL is to parameterize the policy and then iteratively optimize the policy over the parameter space. Let $\{\pi_w : \mathcal{S} \to \mathcal{P}(\mathcal{A}) \mid w \in \mathcal{W}\}$ be a parameterized policy class, where $\mathcal{W}$ is the parameter space. Then, the problem in (2) can be written as

$$\max_{w \in \mathcal{W}} \boldsymbol{F}(\pi_w), \text{ s.t. } f_i(\pi_w) \le c_i, \ \forall i = m+1, \ldots, m+p.$$

In CMORL, we extend the notion of the Pareto frontier, which is defined to compare the policies, from the unconstrained MDP [39] to the safety-constrained MDP.

Definition 3.1 (Safe Pareto Frontier). For any two policies $\pi, \pi' \in \Pi$, we say that π dominates π' if $f_i(\pi) \le f_i(\pi')$ for all i, and there exists one i such that $f_i(\pi) < f_i(\pi')$; otherwise, we say that π does not dominate π'. A solution $\pi^* \in \Pi_{\text{safe}}$ is called safe Pareto optimal if it is not dominated by any other safe policy in Π_{safe}. The set of all safe Pareto optimal policies is the safe Pareto frontier.

In this paper, we assume that there exists at least a safe Pareto optimal policy for the problem in (2). The primary aim of CMORL is to identify a safe Pareto optimal policy. Nevertheless, the concurrent learning of multiple objectives introduces a complex optimization problem, as it entails the consideration of numerous objectives simultaneously. This complexity arises from the need to effectively balance trade-offs between conflicting objectives while maintaining safety constraints throughout the learning process. [8]. The most popular multi-objective formulation in practice is the linear scalarization of all objectives given relative preferences for each objective $\xi_i, i \in [m]$:

$$\max_{w \in \mathcal{W}} \boldsymbol{\xi}^\top \boldsymbol{F}(\pi_w), \text{ s.t. } f_i(\pi_w) \le c_i, \ \forall i = m+1, \ldots, m+p.$$

Even when this linear scalarization formulation gives exactly the true objective, directly optimizing it could lead to undesirable performance due to conflicting gradients, dominating gradients, and high curvature [9].

In this paper, we aim to find a safe Pareto optimal solution using the gradient-based method by starting from an arbitrary initialization policy π_t and iteratively finding the next policy π_{t+1} by moving against a direction $\boldsymbol{d}_t$ with step size η_t, i.e., $\pi_{t+1} = \pi_t + \eta_t \boldsymbol{d}_t$. The design of the direction $\boldsymbol{d}_t$ is the key to the success of CMORL. A good direction $\boldsymbol{d}_t$ should enable us to move from a policy π_t to π_{t+1} such that either π_{t+1} dominates π_t or π_{t+1} improves the hard constraint satisfaction compared with π_t, or both.

IV. Constraint-Rectified Multi-Objective Policy Optimization (CR-MOPO)

In this section, we introduce a general framework called CR-MOPO which decomposes safe Pareto optimal policy learning into three sub-problems and iterates until convergence:

1) Policy evaluation: estimate Q-functions given the current policy.
2) Policy improvement for the multi-objectives: update policy based on the manipulated NPG of multi-objectives when constraints are all approximately satisfied.
3) Constraint rectification: update policy based on the NPG of an unsatisfied constraint when constraints are not all approximately satisfied.

Algorithm 1 summarizes this three-step constrained multi-objective policy improvement framework and Algorithm 2 provides a concrete realization with our novel conflict-averse NPG method. Note, based on our theoretical guarantee on the time-average convergence, policy π_{out} can be uniformly chosen from $\mathcal{N}_0$, the detail proof is provided in Section VIII. To ease the presentation and better illustrate the main idea, we will focus on the tabular MDP setting in this section. The extension to the more practical setting of deep RL will be discussed in Section VIII.

A. Policy Evaluation

In this step, we aim to learn Q-functions that can effectively evaluate the preceding policy π_t. To achieve this, we train individual Q-functions for each objective and constraint. In principle, any Q-learning algorithm can be used, as long as the target Q-value is computed with respect to π_t.

a) Temporal difference (TD) learning: In TD learning, each iteration takes the form of

$$Q_{i,k+1}^{\pi_w}(s,a) = Q_{i,k}^{\pi_w} + \ell_k \left[r_i(s,a) + \gamma Q_{i,k}^{\pi_w}(s',a') - Q_{i,k}^{\pi_w}(s,a) \right], \quad (3)$$

Algorithm 1: CR-MOPO: Constraint-Rectified Multi-Objective Policy Optimization Framework.

```
Inputs: initial parameter $\pi_{w_0}$, empty set $\mathcal{N}_0$.
for $t = 0, \ldots, T-1$ do
    Policy evaluation under $\pi_{w_t}$ for all objectives and constraints.
    if constraints are all satisfied then
        Add $\pi_{w_t}$ into set $\mathcal{N}_0$.
        Compute the multi-objective policy update direction $\boldsymbol{d}$ and update policy using $\boldsymbol{d}$.
    else
        Choose any unsatisfied constraint $i_t$ and update policy towards minimize $f_{i_t}(\pi_{w_t})$.
    end if
end for
Outputs: $\pi_{\text{out}}$ uniformly chosen from $\mathcal{N}_0$.
```

where $s \sim \mu_{\pi_w}, a \sim \pi_w(s), s' \sim \mathrm{P}(\cdot \mid s, a), a' \sim \pi_w(s')$, and ℓ_k is the learning rate. It has been shown in [40], [41] that the iteration in (3) converges to the fixed point which is the state-action value $Q_i^{\pi_w}$. After performing K_{TD} iterations of (3), we let the estimation $\bar{Q}_i(s,a) = Q_{i,K_{\text{TD}}}^{\pi_w}(s,a)$.

b) Unbiased Q-estimation: To obtain an unbiased estimation of the state-action value [42], we can perform Monte-Carlo rollouts for a trajectory with the horizon $H \sim \mathrm{Geom}(1-\gamma^{1/2})$, where $\mathrm{Geom}(x)$ denotes a geometric distribution with parameter x, and estimate the state-action value function along the trajectory $(s_0, a_0, \ldots, s_H, a_H)$ as follows:

$$\bar{Q}_i(s_0, a_0) = r_i(s_0, a_0) + \sum_{h=1}^{H} \gamma^{h/2} r_i(s_h, a_h). \tag{4}$$

B. Policy Improvement for Multi-Objectives

1) Conflict-Averse Natural Policy Gradient (CA-NPG): The policy gradient [43] of the value function $f_i(\pi_w)$ has been derived as $\nabla f_i(\pi_w) = \mathbb{E}[Q_i^{\pi_w}(s,a)\phi_w(s,a)]$, where $\phi_w(s,a) := \nabla_w \log \pi_w(a \mid s)$ is the score function. However, the standard policy gradient does not effectively reflect the statistical manifold (the family of probability distributions that represents the policy function) that the policy operates on. To prevent the policy itself from changing too much during an update, we need to consider how sensitive the policy is to parameter changes.

Thus, in the multi-objectives policy optimization, we aim to choose an update direction $\boldsymbol{d}$ to increase every individual value function while imposing the constraint on the allowed changes of an update in terms of the KL divergence of the policy. To do so, we consider the following constrained optimization problem:

$$\max_{\boldsymbol{d}: D_{\text{KL}}(\pi_w | \pi_{w+\boldsymbol{d}}) \leq \epsilon_0} \min_{i \in [m]} \left\{ \xi_i \left(f_i(w + \boldsymbol{d}) - f_i(w) \right) \right\} \tag{5}$$

where ϵ_0 is the pre-defined threshold for allowed policy changes. By using the first-order Taylor approximation for the value improvement, the second-order Taylor approximation for the KL divergence constraint and the Lagrangian relaxation, the problem (5) can be rewritten as

$$\max_{\boldsymbol{d}} \min_{i \in [m]} \left\{ \xi_i \nabla f_i(w)^\top \boldsymbol{d} - \frac{\psi_1}{2} \boldsymbol{d}^\top \tilde{F}(w) \boldsymbol{d} \right\}, \tag{6}$$

where $\psi_1 > 0$ is a pre-specified hyper-parameter to control the allowed changes in policy space and $\tilde{F}(w)$ is the Fisher information matrix defined as $\tilde{F}(w) = \nabla^2_{w'} D_{\text{KL}}(\pi_w \mid \pi_{w'})|_{w'=w} = \mathbb{E}_{\nu_{\pi_w}}[\phi_w(s,a)\phi_w(s,a)^\top]$. For a single objective f_i, the solution of (6) leads to the well-known NPG update [44] which is defined as $\tilde{F}(w)^\dagger \nabla f_i(\pi_w)$. Note that TRPO [45] can be viewed as the NPG approach with adaptive stepsize.

With the above problem formulation, we aim to find an update direction that minimizes the gradient conflicts. The gradient conflict refers to the case when a selected gradient step, say $\boldsymbol{d}$, conflicts with some individual gradient in the multi-objective optimization: $\exists i, \boldsymbol{d}_i^\top \boldsymbol{d} < 0$. However, there always exists a gradient step that does not conflict with other gradients, e.g. zero vector. Furthermore, inspired by the recent advances in gradient manipulation method [34] which looks for the best update direction within a local ball centered at the weighted averaged gradient, we also constraint search region for the common direction as a circle around the weighted average policy gradient $\boldsymbol{v}_0 = \sum_{i=1}^m \xi_i \nabla f_i(w)$. This yields **Conflict-Averse Natural Policy Gradient (CA-NPG)** which determines the update direction $\boldsymbol{d}$ by solving the following optimization problem

$$\max_{\boldsymbol{d}} \min_{i \in [m]} \left\{ \xi_i \nabla f_i(w)^\top \boldsymbol{d} - \frac{\psi_1}{2} \boldsymbol{d}^\top \tilde{F}(w) \boldsymbol{d} - \frac{\psi_2}{2} \|\boldsymbol{d} - \boldsymbol{v}_0\|^2 \right\}, \tag{7}$$

where $\psi_2 > 0$ is a pre-specified hyper-parameter that controls the deviation from the weighted average policy gradient $\boldsymbol{v}_0$. Furthermore, notice that $\min_i \xi_i \nabla f_i(w)^\top \boldsymbol{d} = \min_{\boldsymbol{\theta} \in S_m} \sum_{i \in [m]} \theta_i \xi_i \nabla f_i(w)^\top \boldsymbol{d}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ and $S_m = \{\boldsymbol{\theta} : \sum_{i=1}^m \theta_i = 1, \theta_i \geq 0\}$. Denote $\nabla f_{\boldsymbol{\theta}}(w) = \sum_{i \in [m]} \theta_i \xi_i \nabla f_i(w)$. The objective in (7) can be written as

$$\max_{\boldsymbol{d}} \min_{\boldsymbol{\theta} \in S_m} \left\{ \nabla f_{\boldsymbol{\theta}}(w)^\top \boldsymbol{d} - \frac{\psi_1}{2} \boldsymbol{d}^\top \tilde{F}(w) \boldsymbol{d} - \frac{\psi_2}{2} \|\boldsymbol{d} - \boldsymbol{v}_0\|^2 \right\}.$$

Since the above objective is concave with respect to $\boldsymbol{d}$ and linear with respect to $\boldsymbol{\theta}$, by switching the min and max, we reach the dual form without changing the solution:

$$\min_{\boldsymbol{\theta} \in S_m} \max_{\boldsymbol{d}} \left\{ \nabla f_{\boldsymbol{\theta}}(w)^\top \boldsymbol{d} - \frac{\psi_1}{2} \boldsymbol{d}^\top \tilde{F}(w) \boldsymbol{d} - \frac{\psi_2}{2} \|\boldsymbol{d} - \boldsymbol{v}_0\|^2 \right\}.$$

After a few steps of calculus (details are in the appendix, available online), we derive the following optimization problem with respect to the variable $\boldsymbol{\theta}$:

$$\begin{aligned} \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in S_m} & \nabla f_{\boldsymbol{\theta}}^\top \left(\psi_1 \tilde{F} + \psi_2 I \right)^{-1} (\nabla f_{\boldsymbol{\theta}} + \psi_2 \boldsymbol{v}_0) \\ & - \frac{\psi_1}{2} (\nabla f_{\boldsymbol{\theta}} + \psi_2 \boldsymbol{v}_0)^\top \left(\psi_1 \tilde{F} + \psi_2 I \right)^{-1} \\ & \tilde{F} \left(\psi_1 \tilde{F} + \psi_2 I \right)^{-1} (\nabla f_{\boldsymbol{\theta}} + \psi_2 \boldsymbol{v}_0) \\ & - \frac{\psi_2}{2} \left\| \left(\psi_1 \tilde{F} + \psi_2 I \right)^{-1} (\nabla f_{\boldsymbol{\theta}} + \psi_2 \boldsymbol{v}_0) \right\|^2, \end{aligned}$$

and the optimal update direction is given by

$$\boldsymbol{d}^* := \boldsymbol{\lambda}^\top \nabla \boldsymbol{F}^\psi = \sum_{i \in [m]} \lambda^i \left(\psi_1 \tilde{F} + \psi_2 I\right)^{-1} \nabla f^i, \tag{8}$$

where $\lambda^i = \xi_i(\theta_i^* + \psi_2)$ and $\nabla \boldsymbol{F}_i^\psi = (\psi_1 \tilde{F} + \psi_2 I)^{-1} \nabla f^i$. To simplify the notation, we omit the superscript ψ in $\nabla \boldsymbol{F}^\psi$ for the subsequent sections.

2) Correlation-Reduction for Stochastic Gradient Manipulation: In practice, we only obtain noisy policy gradient feedback $\widehat{\nabla \boldsymbol{F}}(w_t)$, where the stochastic noise is due to the finite sampled trajectories for the estimation of $Q_i^{\pi_{w_t}}$. It has been shown in [39] that the gradient manipulation methods may fail to converge to a Pareto optimal solution under the stochastic setting. This convergence gap is mainly caused by the strong correlation between the weights $\boldsymbol{\lambda}_t$ and the stochastic gradients $\widehat{\nabla \boldsymbol{F}}(w_t)$ which yields a biased composite gradient. To address this issue in CA-NPG, we consider two conditions. The first is that the NPG estimator variance asymptotically converges to 0. For example, this can be achieved by estimating $Q_i^{\pi_{w_t}}$ using TD learning in (3) with sufficiently large K_{TD}. The second is to reduce the variances of $\boldsymbol{\lambda}_\tau$ by adopting a momentum mechanism [39] with coefficient α_t on the update of composite weights

$$\widehat{\boldsymbol{\lambda}}_\tau = \alpha_\tau \widehat{\boldsymbol{\lambda}}_{\tau-1} + (1 - \alpha_\tau) \boldsymbol{\lambda}_\tau, \tag{9}$$

where $\boldsymbol{\lambda}_\tau$ is computed by CA-NPG algorithms.

C. Constraint Rectification

We then check whether there exists a hard constraint $i \in \{m+1, \ldots, m+p\}$ such that the (approximated) constraint function violates the condition. If so, we take one-step update of the policy using NPG towards minimizing the corresponding constraint function $f_i(\pi_{w_t})$ to enforce the constraint:

$$w_{t+1} = w_t - \eta \tilde{F}(w)^\dagger \nabla f_i \left(\pi_w\right).$$

If multiple constraints are violated, we can choose to minimize any one of them. Otherwise, we take one update of the policy towards maximizing the multi-objectives.

D. Comparison With Learning Preferences and Policies in Parallel (LP3) [11]

Compared with SOTA safe multi-objective RL method, LP3, our new framework is different in both multi-objective optimization and hard constraint satisfaction. First, LP3 chooses MO-MPO [32] as the multi-objective optimizer which encodes the objective preferences in a scale-invariant way through the allowed KL divergence for the updated policy using each objective. On the other hand, our multi-objective optimization method is based on linear scalarization coupled with novel NPG manipulation which encodes the preference in a more straightforward way and is tailored to RL to address the conflicting gradients and dominating gradients. Second, LP3 can be regarded as a primal-dual approach where the additional dual variables are introduced as the adaptive weights for the constraints. This relaxes hard constraints in safe multi-objective RL problems to new objectives where the associated weights are adjusted based on the constraint violation conditions. On the other hand, our primal-based method does not suffer from extra hyperparameter tuning and dual update and can be implemented as easily as unconstrained policy optimization algorithms.

Algorithm 2: CR-MOPO With CA-NPG as Multi-Objective Optimizer.

```
Inputs: initial parameter $w_0$, empty set $\mathcal{N}_0$, $\tau = 0$.
for $t = 0, \ldots, T-1$ do
  Policy evaluation under $\pi_{w_t}$ : $\bar{Q}_i^t(s,a) \approx Q_i^{\pi_{w_t}}(s,a)$
  for all $i = 1, \ldots, m+p$.
  Collect pairs $(s^j, a^j) \in \mathcal{B}_t \sim \rho \cdot \pi_{w_t}$, compute
  constrain estimation $\bar{J}_{i,\mathcal{B}_t} = \sum_{j \in \mathcal{B}_t} \frac{1}{|\mathcal{B}_t|} \bar{Q}_t^i(s^j, a^j)$
  for all $i = 1, \ldots, m+p$, where $j$ is the index for the
  sampled pairs in $\mathcal{B}_t$.
  if $\bar{J}_{i,\mathcal{B}_t} \leq c_i + \beta$ for all $i = m+1, \ldots, m+p$ then
    $\tau \leftarrow \tau + 1$; add $w_t$ into set $\mathcal{N}_0$.
    Compute the weights $\boldsymbol{\lambda}_\tau$ using (8) and reduce the
    correlation by (9).
    Compute the multi-objective policy gradient
    $\boldsymbol{d}_\tau = \widehat{\boldsymbol{\lambda}}_\tau^\top \widehat{\nabla \boldsymbol{F}}(w_t)$.
    Take one-step policy update: $w_{t+1} = w_t + \eta \boldsymbol{d}_\tau$.
  else
    Choose any $i_t \in \{m+1, \ldots, m+p\}$ such that
    $\bar{J}_{i_t,\mathcal{B}_t} > c_{i_t} + \beta$.
    Take one-step policy update towards minimize
    $J_{i_t}(w_t)$: $w_{t+1} \leftarrow w_t - \eta \tilde{F}(w_t)^\dagger \nabla f_i(w_t)$.
  end if
end for
Outputs: $w_{\text{out}}$ uniformly chosen from $\mathcal{N}_0$.
```

V. Theoretical Analysis

In this section, we establish the convergence and the constraint violation guarantee for CR-MOPO in the tabular settings under the softmax parameterization and CA-NPG. In the tabular setting, we consider the softmax parameterization. For any $w \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, the corresponding softmax policy π_w is defined as $\pi_w(a \mid s) := \frac{\exp(w(s,a))}{\sum_{a' \in \mathcal{A}} \exp(w(s,a'))}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Clearly, the policy class defined above is complete, as any stochastic policy in the tabular setting can be represented in this class. Since the adaptive weights $\boldsymbol{\lambda}_t$ of CA-NPG may not be constrained in the probability simplex. Hence, we consider the following mild assumption on the boundedness of $\boldsymbol{\lambda}_t$. In the following, we use the subscript t for update steps, and superscript i for i-th objective. When the timestep t can be understood from the context, we only use the objective index.

Assumption 5.1. For the CA-NPG mechanism, there exists finite constants $B_1 > 0$ and $B_2 > 0$ such that $0 \leq \lambda_t^i \leq B_1, \sum_{i=1}^m \lambda_t^i \geq B_2$ for all $t = 1, \ldots, T$, $i = 1, \ldots, m$.

Based on the definition of $\boldsymbol{\lambda}_t$, if we assume that relative preferences $\{\xi_i\}_{i=1}^m \in \mathcal{S}_m$ for all i. Then, we have $0 \leq \lambda_t^i \leq 1 + \psi_2$ and $\sum_i \lambda_t^i \geq \psi_2$. Thus, we can take $B_1 = 1 + \psi_2$ and $B_2 = \psi_2$, which makes Assumption 5.1 holds.

For multi-objective optimization, if there *exists* $\boldsymbol{\lambda}^* \in S_m$ such that $w^* = \arg\min_w \boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi_w)$, then w^* is (weak) Pareto optimal [Theorem 5.13 and Lemma 5.14 in [46]]. Thus, we use $\min_{\boldsymbol{\lambda}^* \in S_m}(\boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi^*) - \boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi_{w_{\text{out}}}))$ to measure the convergence to a Pareto optimal policy where minimization operator of $\boldsymbol{\lambda}^*$ is from the existence condition. The following theorem characterizes the convergence rate of Algorithm 2 in terms of the Pareto optimal policy convergence and hard constraint violations. The proof can be found in the Appendix VIII, available online.

Theorem 5.2. Consider Algorithm 2 in the tabular setting with softmax policy parameterization and any policy initialization $w_0 \in \mathcal{R}^{|\mathcal{S}||\mathcal{A}|}$. Let the tolerance be $\beta = \mathcal{O}(\frac{mB_1\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2\sqrt{T}})$ and the learning rate for the CA-NPG and NPG be $\eta = \mathcal{O}(\frac{(1-\gamma)^2}{mB_1\sqrt{|\mathcal{S}||\mathcal{A}|T}})$. Depending on the choice of the state-action value estimator, the following holds.

- If TD-learning in (3) is used for policy evaluation with $K_{\text{TD}} = \widetilde{\mathcal{O}}((\frac{T}{(1-\gamma)^2|\mathcal{S}||\mathcal{A}|})^{\frac{1}{\sigma}})$, $\ell_k = \mathcal{O}(\frac{1}{k^\sigma})$ and $\alpha_\tau = 0$ for $0 < \sigma < 1$, then with probability $1-\delta$, we have

$$\mathbb{E}\left[\min_{\boldsymbol{\lambda}^* \in S_m}\left(\boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi^*) - \boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi_{w_{\text{out}}})\right)\right] \le \frac{\beta}{B_2},$$
$$\mathbb{E}[f_i(\pi_{w_{\text{out}}})] - c_i \le \beta,$$

  for all $i = \{m+1, \ldots, m+p\}$, where the expectation is taken only with respect to selecting w_{out} from $\mathcal{N}_0$.

- If unbiased Q-estimation in (4) is used for policy evaluation with $\alpha_\tau \ge 1 - \frac{1-\gamma}{m\tau\sqrt{|\mathcal{S}||\mathcal{A}|}}$, we have

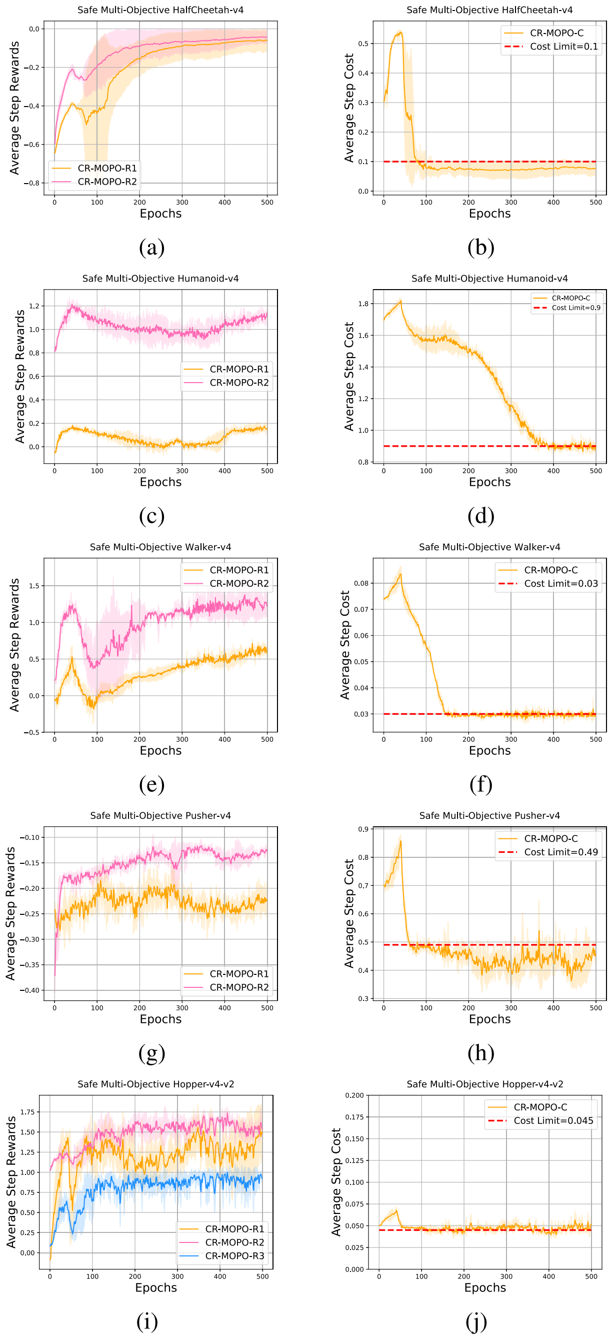
$$\mathbb{E}\left[\min_{\boldsymbol{\lambda}^* \in S_m}\left(\boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi^*) - \boldsymbol{\lambda}^{*\top}\boldsymbol{F}(\pi_{w_{\text{out}}})\right)\right] \le \frac{\beta}{B_2},$$
$$\mathbb{E}[f_i(\pi_{w_{\text{out}}})] - c_i \le \beta,$$

  for all $i = \{m+1, \ldots, m+p\}$, where the expectation is taken with respect to selecting w_{out} from $\mathcal{N}_0$ and the randomness of $Q^i_{\pi_{w_t}}$ estimation.

As shown in Theorem 5.2, our method is guaranteed to find a safe Pareto optimal policy under some mild conditions while there is no convergence guarantee for LP3 [11]. Furthermore, results for unbiased Q-estimation imply that the correlation reduction mechanism could help the convergence even if we do not have an asymptotically increasing trajectory for policy evaluation, such as $K_{\text{TD}} = \widetilde{\mathcal{O}}(T^{1/\sigma})$ in TD-learning.

VI. Experiments

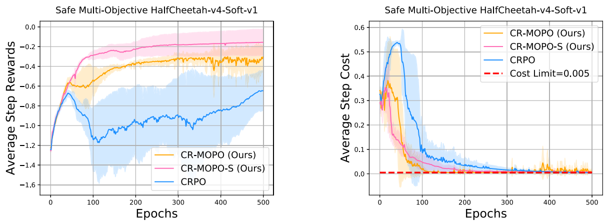
Environment Settings: We have designed a benchmark, referred to as *Safe Multi-Objective MuJoCo*, to evaluate our algorithms within the MuJoCo framework [12], [47]. A comprehensive description of this benchmark is provided in Appendix XI-A, available online, where we introduce environments such as Safe Multi-Objective HalfCheetah, Safe Multi-Objective Hopper, Safe Multi-Objective Humanoid, Safe Multi-Objective Swimmer, Safe Multi-Objective Walker, and Safe Multi-Objective Pusher to examine the effectiveness of our proposed methods.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

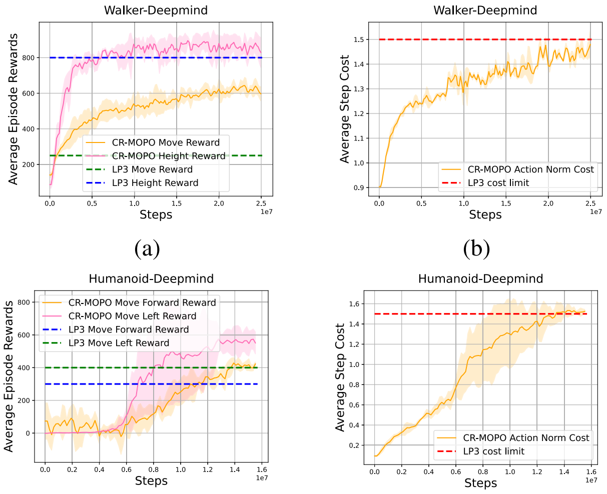
Fig. 1. CR-MOPO on Safe Multi-Objective MuJoCo environments regarding the reward and safety performance.

CR-MOPO Exhibits Strong Performance in Challenging Safe Multi-Objective Environments: As illustrated in Fig. 1, a series of experiments are conducted across various challenging tasks. The cost limits for each step were set as follows: HalfCheetah-v4 at 0.1, Humanoid-v4 at 0.9, Walker-v4 at 0.03, Pusher-v4 at 0.49, and Hopper-v4-v2 at 0.045. The optimization of safety violations is performed after 40 Epochs for all tasks, with the exception of the Humanoid-dm task, for which the optimization is carried out after 5 Epochs. The experimental results demonstrate that our method is capable of ensuring monotonic improvement in each task's reward while maintaining safety



(a) (b)

Fig. 2. (a) and (b) show the comparison results in terms of CR-MOPO, CR-MOPO-S and CRPO [13] on a Safe Multi-Objective MuJoCo environment, Safe Multi-Objective HalfCheetah, the cost limit is 0.005, we start to optimize safety violation after 40 Epochs.



(a) (b)

(c) (d)

Fig. 3. Compared with LP3 [11], on Safe Multi-Objective Walker-dm and Safe Multi-Objective Humanoid-dm environments.

across all challenging tasks. For further experimental details, refer to Appendix XI-A, available online.

Exploiting Constraints' Dual Nature Elevates Both Safety and Performance: In this study, we investigate *the implications of integrating a safety constraint both as an independent constraint and as an auxiliary objective within a multi-objective framework.*, referred to as CR-MOPO-Soft (CR-MOPO-S). The pseudocode for CR-MOPO-S is shown in Appendix X, available online, Algorithm 3. Within this framework, the constraint is seamlessly integrated by allocating a specified weight to performance, for instance, a weight of 1.0.

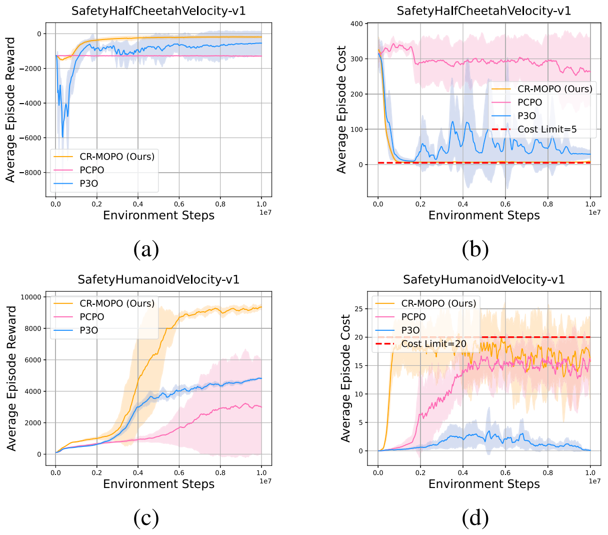
We consider that treating the constraint function as an objective can effectively buffer the feasible set's boundary, facilitating navigation toward a "deep safe" set. This, in turn, ensures uninterrupted progress in performance. This mechanism is crucial in heavily constrained systems, where operating near the safety boundary can lead to constraint violations and unstable behaviors, as seen in other safe learning approaches such as CRPO [13], CPO [28], and PCPO [29].

Regarding the effectiveness of CR-MOPO-S, we expand the scope of our experiments and scenarios, focusing on a straightforward Constrained Markov Decision Process (CMDP) framework. In this setup, the primary emphasis is on reward optimization for a single objective, while incorporating safety constraints. We compare CR-MOPO-S against CR-MOPO, which excludes the constraint from its objective, and benchmark it against the state-of-the-art safe RL algorithm, CRPO [13]. CRPO is an important safe RL algorithm that has consistently demonstrated its effectiveness in terms of reward and safety performance, outperforming other safe RL approaches such as PDO [48]. Since CRPO is specifically designed for single-objective safe RL, we aggregate multiple objectives into a single objective to ensure compatibility with the algorithm. Detailed implementation specifics can be found in Appendix XI-C, available online.

As illustrated in Fig. 2, our introduced algorithms, CR-MOPO and CR-MOPO-S, surpass CRPO in both reward and safety performance. Notably, CR-MOPO-S showcases superior constraint adherence and performance optimization compared to CR-MOPO. This can be intuitively understood by comparing it to running on a road: if one runs too close to the edge, even a slight misstep may require repositioning, potentially sacrificing speed. In contrast, running closer to the center ensures both safety and optimal speed. Algorithmically, while constraints are typically treated as binary outcomes (satisfied or not), objectives pursue a continuous path of improvement. By softening the binary nature of constraints, our approach effectively navigates complex scenarios, avoiding the frequent toggling at the boundary of the feasible set.

Benefit of Boundary-Aware Policy Learning (in Comparison to LP3 [11]): In this section, we evaluate our algorithm against LP3 [11]. LP3 optimizes policies under constraints by prioritizing objectives based on preference values. This approach has demonstrated success in tackling complex tasks, such as Humanoid-dm and Walker-dm, from the DeepMind Control Suite [49].

As evidenced in Fig. 3, our algorithm consistently improves over LP3 [11] while ensuring safety. Specifically, in the Walker-dm task (refer to Fig. 3(a) and (b)), with a cost limit of 1.5, our method achieves move and height rewards surpassing 600 and 800 respectively, while LP3 manages roughly 250 and 800 in the same measures. Regarding the Humanoid-dm task, while LP3 scores around 400 and 300 for the move left and move forward rewards with a cost limit of 1.5, our approach consistently reaches approximately 600 for move left and at least 400 for move forward rewards. These empirical results highlight the clear advantage of our algorithm over LP3 [11]. This superiority stems from our method's unique feature of dynamically adapting the policy learning strategy based on handling constraint boundaries. Within the safety set, our algorithm strategically *eases* constraint satisfaction, reducing the risk of conflicting gradient updates that often occur near boundaries. This boundary-aware learning approach contrasts with LP3's methodology, which translates constraints into objectives through predefined preferences. While effective in certain scenarios, LP3's strategy fails to fully exploit the potential of the safety set. Although our approach may occasionally favor policies near the safety boundary, this tendency can be effectively mitigated by selecting stricter constraint thresholds.



(a) (b) (c) (d)

Fig. 4. (a), (b), (c), and (d) show the experimental results compared our method with PCPO [29] and P3O [51] on *Omnisafe* tasks, e.g., *SafetyHalfCheetahVelocity-v1* and *SafetyHumanoidVelocity-v1* tasks. Our method performs better than the strong baselines regarding safety and reward performance.

Comparison With Safe RL Methods: To evaluate the effectiveness of our approach in safe RL tasks, we implement our algorithm on the widely popular safe RL benchmark, *Omnisafe* [50]. This implementation facilitated comparative analyses with several SOTA baselines, including Projection-based Constrained Policy Optimization (PCPO) [29] and Penalized Proximal Policy Optimization (P3O) [51]. All experiments are conducted in the same environment settings. For detailed experiment settings, see the Appendix XI-B, available online.

The comparative results, as illustrated in Fig. 4(a) and (b), demonstrate the performance of our algorithm on the *SafetyHalfCheetahVelocity-v1* task. Notably, our method exhibits superior performance relative to the SOTA baselines. A significant observation is the underperformance of PCPO in this task, both in terms of reward and safety performance. While P3O performs better than PCPO regarding safety performance, it violates safety constraints. In contrast, our method consistently ensured safety, surpassing the performance of the SOTA baselines.

Further experiments, depicted in Fig. 4(c) and (d), are conducted on the *SafetyHumanoidVelocity-v1* task. Again, our method surpasses the SOTA baselines regarding reward and safety performance. This consistent superiority across various tasks highlights our algorithm's effectiveness within the safe RL domain.

VII. Conclusion

In this study, we aim to handle a balance between the performance of individual objectives while maintaining safety in a multi-objective RL context. To this end, a primal-based safe multi-objective RL framework is proposed, which resolves multiple conflict gradients through gradient manipulation and employs constraint rectification to identify safety policies during multi-objective optimization. Moreover, the analysis of convergence and safety violations are provided. In conclusion, we deploy our practical algorithms on several challenging, safe multi-objective RL environments and compare our method with the SOTA safe RL baselines and safe multi-objective RL algorithms. The experiment results indicate that our method can perform better than SOTA-safe RL baselines and SOTA-safe multi-objective RL algorithms regarding the balance between each objective performance and safety violation. In the future, we plan to deploy our algorithm in real-world scenarios and try to leverage the foundation models [52] with our method to address safe multi-objective RL robustness problems.

References

[1] S. Gu, A. Kshirsagar, Y. Du, G. Chen, J. Peters, and A. Knoll, "A human-centered safe robot reinforcement learning framework with interactive behaviors," *Front. Neurorobot.*, vol. 17, 2023, Art. no. 1280341.

[2] S. Gu et al., "Safe multi-agent reinforcement learning for multi-robot control," *Artif. Intell.*, vol. 319, 2023, Art. no. 103905.

[3] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.

[4] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.

[5] S. Gu, G. Chen, L. Zhang, J. Hou, Y. Hu, and A. Knoll, "Constrained reinforcement learning for vehicle motion planning with topological reachability analysis," *Robotics*, vol. 11, no. 4, 2022, Art. no. 81.

[6] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 27 730–27 744.

[7] A. Charpentier, R. Elie, and C. Remlinger, "Reinforcement learning in economics and finance," *Comput. Econ.*, vol. 62, pp. 425–462, 2023.

[8] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022.

[9] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 5824–5836.

[10] V. Kurin, A. De Palma, I. Kostrikov, S. Whiteson, and P. K. Mudigonda, "In defense of the unitary scalarization for deep multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 12 169–12 183.

[11] S. Huang et al., "A constrained multi-objective reinforcement learning framework," in *Proc. Conf. Robot Learn.*, PMLR, 2022, pp. 883–893.

[12] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. 2012 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 5026–5033.

[13] T. Xu, Y. Liang, and G. Lan, "CRPO: A new approach for safe reinforcement learning with convergence guarantee," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 11 480–11 491.

[14] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 4431–4506, 2021.

[15] S. Gu et al., "A review of safe reinforcement learning: Methods, theory and applications," 2022, *arXiv:2205.10330*.

[16] R. Wu, Y. Zhang, Z. Yang, and Z. Wang, "Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 25 439–25 451.

[17] N. Vithayathil Varghese and Q. H. Mahmoud, "A survey of multi-task deep reinforcement learning," *Electronics*, vol. 9, no. 9, 2020, Art. no. 1363.

[18] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration in finite Markov decision processes with Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4312–4320.

[19] F. Berkenkamp and A. P. Schoellig, "Safe and robust learning control with Gaussian processes," in *Proc. IEEE 2015 Eur. Control Conf.*, 2015, pp. 2496–2501.

[20] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 997–1005.

[21] A. Wachi, Y. Sui, Y. Yue, and M. Ono, "Safe exploration and optimization of constrained MDPs using Gaussian processes," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6548–6555.

[22] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8103–8112.

[23] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based safe policy optimization for continuous control," 2019, *arXiv:1901.10031*.

[24] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *Proc. 2018 IEEE Conf. Decis. Control*, 2018, pp. 6059–6066.

[25] X. Li and C. Belta, "Temporal logic guided safe reinforcement learning using control barrier functions," 2019, *arXiv:1903.09885*.

[26] Z. Marvi and B. Kiumarsi, "Safe reinforcement learning: A control barrier function optimization approach," *Int. J. Robust Nonlinear Control*, vol. 31, no. 6, pp. 1923–1940, 2021.

[27] N. Fulton and A. Platzer, "Safe reinforcement learning via formal methods: Toward safe control through proof and learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6485–6492.

[28] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2017, pp. 22–31.

[29] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," in *Proc. Int. Conf. Learn. Representations*, 2020.

[30] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *Proc. Int. Conf. Learn. Representations*, 2018.

[31] D. Ying, M. A. Guo, Y. Ding, J. Lavaei, and Z.-J. Shen, "Policy-based primal-dual methods for convex constrained markov decision processes," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 10 963–10 971.

[32] A. Abdolmaleki et al., "A distributional view on multi-objective policy optimization," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 11–22.

[33] A. Navon et al., "Multi-task learning as a bargaining game," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 16 428–16 446.

[34] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, "Conflict-averse gradient descent for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18 878–18 890.

[35] J.-A. Désidéri, "Multiple-gradient descent algorithm (MGDA) for multiobjective optimization," *Comptes Rendus Mathematique*, vol. 350, no. 5-6, pp. 313–318, 2012.

[36] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, "Empirical evaluation methods for multiobjective reinforcement learning algorithms," *Mach. Learn.*, vol. 84, no. 1-2, pp. 51–80, 2011.

[37] S. Parisi, M. Pirotta, and M. Restelli, "Multi-objective reinforcement learning through continuous Pareto manifold approximation," *J. Artif. Intell. Res.*, vol. 57, pp. 187–227, 2016.

[38] S. Parisi, M. Pirotta, and J. Peters, "Manifold-based multi-objective policy search with sample reuse," *Neurocomputing*, vol. 263, pp. 3–14, 2017.

[39] S. Zhou, W. Zhang, J. Jiang, W. Zhong, J. Gu, and W. Zhu, "On the convergence of stochastic multi-objective gradient manipulation and beyond," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 38 103–38 115.

[40] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Proc. Conf. Learn. Theory*, PMLR, 2018, pp. 1691–1692.

[41] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, "Finite sample analyses for TD (0) with function approximation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6144–6160.

[42] K. Zhang, A. Koppel, H. Zhu, and T. Basar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM J. Control Optim.*, vol. 58, no. 6, pp. 3586–3612, 2020.

[43] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 1057–1063.

[44] S. M. Kakade, "A natural policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 1531–1538.

[45] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 1889–1897.

[46] J. John, *Vector Optimization, Theory, Application, and Extensions*. Berlin, Germany: Springer, 2004.

[47] G. Brockman et al., "OpenAI gym," 2016, *arXiv:1606.01540*.

[48] A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," 2019, *arXiv:1910.01708*.

[49] Y. Tassa et al., "Deepmind control suite," 2018, *arXiv:1801.00690*.

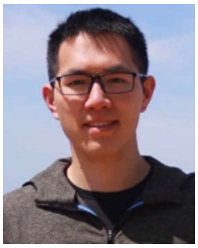
[50] J. Ji et al., "OmniSafe: An infrastructure for accelerating safe reinforcement learning research," 2023, *arXiv:2305.09304*.

[51] L. Zhang et al., "Penalized proximal policy optimization for safe reinforcement learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3744–3750.

[52] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans, "Foundation models for decision making: Problems, methods, and opportunities," 2023, *arXiv:2303.04129*.

Shangding Gu (Graduate Student Member, IEEE) is currently working as a postdoc with UC Berkeley and a guest researcher with the Technical University of Munich. He is one of the organizers of the 1st International Safe Reinforcement Learning Workshop at IEEE MFI 2022. His main research interests include safe learning and planning for robotics and foundation models.

Bilgehan Sel is currently working toward the PhD degree in electrical and computer engineering with Virginia Tech, with his primary research focusing on reinforcement learning, large language models, and optimization, emphasizing safety.

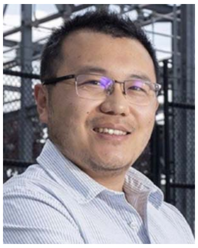


Yuhao Ding received the PhD degree in industrial engineering and operations research from the University of California, Berkeley. He is currently a research analyst with the Cubist Systematic Strategies. His research has been focused on the interdisciplinary problems in control, reinforcement learning, optimization and statistical learning.

Lu Wang received the PhD degree from East China Normal University. She is a researcher with Microsoft Research Asia specializing in reinforcement learning, cloud intelligence, healthcare, finance, and other related areas. She has published more than 20 papers on RL, including ICML, ICLR, and KDD.

Qingwei Lin (Member, IEEE) is a researcher with Microsoft Research Asia and Partner research manager with the DKI (Data, Knowledge, Intelligence) research area. He is leading a team of researchers working on data-driven technologies for cloud intelligence, with innovations in machine learning and data mining algorithms. He got Best Research Paper Award at ISSRE and SIGSOFT Distinguished Paper Award at ESEC/FSE. The research technologies have been transferred into multiple Microsoft product divisions, such as Microsoft Azure, Office365, Windows, Bing, etc, and made substantial improvements. He hosted Microsoft company-wide "Cloud Service Intelligence Summit" as the Chair for five consecutive years.

Alois Knoll (Fellow, IEEE) received the Diploma (MSc) degree in electrical/communications engineering from the University of Stuttgart, Stuttgart, Germany, in 1985, and the PhD degree in computer science from the Technical University of Berlin (TU Berlin), Berlin, Germany, in 1988. He served on the faculty of the Computer Science Department, TU Berlin until 1993. He joined the University of Bielefeld, Bielefeld, Germany, as a full professor and the director of the Research Group Technical Informatics until 2001. Since 2001, he has been a professor with the Department of Informatics, Technical University of Munich (TUM), Munich, Germany. He was also on the board of directors of the Central Institute of Medical Technology, TUM. From 2004 to 2006, he was the executive director of the Institute of Computer Science, TUM. He is the editor-in-chief of the neurorobotics journal of frontiers. His research interests include multi-agent systems, adaptive systems, data fusion, and robotics.



Ming Jin received the doctoral degree from EECS department, University of California, Berkeley in 2017. He is an assistant professor with the Bradley Department of Electrical and Computer Engineering, Virginia Tech. He was a postdoctoral researcher with the Department of Industrial Engineering and Operations Research, University of California. He has received numerous accolades, including the Commonwealth Cyber Initiative Research Innovation Award, Siebel Scholarship, and multiple Best Paper Awards. His group won the first place in the 2021 CityLearn Challenge. His research focuses on trustworthy AI for engineering and science, emphasizing safety and alignment.