Spectral-Risk Safe Reinforcement Learning with Convergence Guarantees

Dohyeong Kim[1] **Taehyun Cho**[1] **Seungyub Han**[1]
Hojun Chung[1] **Kyungjae Lee**[2*] **Songhwai Oh**[1*]
[1]Dep. of Electrical and Computer Engineering, Seoul National University
[2]Dep. of Statistics, Korea University

Abstract

The field of risk-constrained reinforcement learning (RCRL) has been developed to effectively reduce the likelihood of worst-case scenarios by explicitly handling risk-measure-based constraints. However, the nonlinearity of risk measures makes it challenging to achieve convergence and optimality. To overcome the difficulties posed by the nonlinearity, we propose a spectral risk measure-constrained RL algorithm, *spectral-risk-constrained policy optimization (SRCPO)*, a bilevel optimization approach that utilizes the duality of spectral risk measures. In the bilevel optimization structure, the outer problem involves optimizing dual variables derived from the risk measures, while the inner problem involves finding an optimal policy given these dual variables. The proposed method, to the best of our knowledge, is the first to guarantee convergence to an optimum in the tabular setting. Furthermore, the proposed method has been evaluated on continuous control tasks and showed the best performance among other RCRL algorithms satisfying the constraints. Our code is available at https://github.com/rllab-snu/Spectral-Risk-Constrained-RL.

1 Introduction

Safe reinforcement learning (Safe RL) has been extensively researched [2, 15, 25], particularly for its applications in safety-critical areas such as robotic manipulation [3, 17] and legged robot locomotion [28]. In safe RL, constraints are often defined by the expectation of the discounted sum of costs [2, 25], where cost functions are defined to capture safety signals. These expectation-based constraints can leverage existing RL techniques since the value functions for costs follow the Bellman equation. However, they inadequately reduce the likelihood of significant costs, which can result in worst-case outcomes, because expectations do not capture information on the tail distribution. In real-world applications, avoiding such worst-case scenarios is particularly crucial, as they can lead to irrecoverable damage.

In order to avoid worst-case scenarios, various safe RL methods [10, 26] attempt to define constraints using risk measures, such as conditional value at risk (CVaR) [20] and its generalized form, spectral risk measures [1]. Yang et al. [26] and Zhang and Weng [30] have employed a distributional critic to estimate the risk constraints and calculate policy gradients. Zhang et al. [32] and Chow et al. [10] have introduced auxiliary variables to estimate CVaR. These approaches have effectively prevented worst-case outcomes by implementing risk constraints. However, the nonlinearity of risk measures makes it difficult to develop methods that ensure optimality. To the best of our knowledge, even for well-known risk measures such as CVaR, existing methods only guarantee local convergence [10, 32]. This highlights the need for a method that ensures convergence to a global optimum.

*Corresponding authors: kyungjae_lee@korea.ac.kr, songhwai@snu.ac.kr.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In this paper, we introduce a spectral risk measure-constrained RL (RCRL) algorithm called *spectral-risk-constrained policy optimization (SRCPO)*. Spectral risk measures, including CVaR, have a dual form expressed as the expectation over a random variable [7]. Leveraging this duality, we propose a bilevel optimization framework designed to alleviate the challenges caused by the nonlinearity of risk measures. This framework consists of two levels: the *outer problem*, which involves optimizing a dual variable that originates from the dual form of spectral risk measures, and the *inner problem*, which aims to find an optimal policy for a given dual variable. For the inner problem, we define novel risk value functions that exhibit linearity in performance difference and propose a policy gradient method that ensures convergence to an optimal policy in tabular settings. The outer problem is potentially non-convex, so gradient-based optimization techniques are not suitable for global convergence. Instead, we propose a method to model and update a distribution over the dual variable, which also ensures finding an optimal dual variable in tabular settings. As a result, to the best of our knowledge, we present the first algorithm that guarantees convergence to an optimum for the RCRL problem. Moreover, the proposed method has been evaluated on continuous control tasks with both single and multiple constraints, demonstrating the highest reward performance among other RCRL algorithms satisfying constraints in all tasks. Our contributions are summarized as follows:

- We propose a bilevel optimization framework for RCRL, effectively addressing the complexities caused by the nonlinearity of risk measures.
- The proposed method demonstrates convergence and optimality in tabular settings.
- The proposed method has been evaluated on continuous control tasks with both single and multiple constraints, consistently outperforming existing RCRL algorithms.

2 Related Work

RCRL algorithms can be classified according to how constraints are handled and how risks are estimated. There are two main approaches to handling constraints: the Lagrangian [26, 29, 32] and the primal [15, 30] approaches. The Lagrangian approach deals with dual problems by jointly updating dual variables and policies, which can be easily integrated into the existing RL framework. In contrast, the primal approach directly tackles the original problem by constructing a subproblem according to the current policy and solving it. Subsequently, risk estimation can be categorized into three types. The first type [29, 32] uses an auxiliary variable to precisely estimate CVaR through its dual form. Another one [26, 30] approximates risks as the expectation of the state-wise risks calculated using a distributional critic [8]. The third one [15] approximates CVaR by assuming that cost returns follow Gaussians. Despite these diverse methods, they only guarantee convergence to local optima. In the field of risk-sensitive RL, Bäuerle and Glauner [7] have proposed an algorithm for spectral risk measures that ensures convergence to an optimum. They introduced a bilevel optimization approach, like our method, but did not provide a method for solving the outer problem. Furthermore, their approach to the inner problem is value-based, making it complicated to extend to RCRL, which typically requires a policy gradient approach.

3 Backgrounds

3.1 Constrained Markov Decision Processes

A constrained Markov decision process (CMDP) is defined as $\langle S, A, P, \rho, \gamma, R, \{C_i\}_{i=1}^N \rangle$, where S is a state space, A is an action space, P is a transition model, ρ is an initial state distribution, γ is a discount factor, $R : S \times A \times S \mapsto [-R_{\max}, R_{\max}]$ is a reward function, $C_i : S \times A \times S \mapsto [0, C_{\max}]$ is a cost function, and N is the number of cost functions. A policy $\pi : S \mapsto \mathcal{P}(A)$ is defined as a mapping from the state space to the action distribution. Then, the state distribution given π is defined as $d_\rho^\pi(s) := (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}(s_t = s)$, where $s_0 \sim \rho$, $a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$ for $\forall t$. The state-action return of π is defined as:

$$Z_R^\pi(s,a) := \sum_{t=0}^\infty \gamma^t R(s_t, a_t, s_{t+1}) \;\Big|\; s_0 = s,\ a_0 = a,\ a_t \sim \pi(\cdot|s_t),\ s_{t+1} \sim P(\cdot|s_t, a_t)\ \forall t.$$

The state return and the return are defined, respectively, as:

$$Y_R^\pi(s) := Z_R^\pi(s,a) \;\Big|\; a \sim \pi(\cdot|s), \quad G_R^\pi := Y_R^\pi(s) \;\Big|\; s \sim \rho.$$

The returns for a cost, $Z^\pi_{C_i}$, $Y^\pi_{C_i}$, and $G^\pi_{C_i}$, are defined by replacing the reward with the ith cost. Then, a safe RL problem can be defined as maximizing the reward return while keeping the cost returns below predefined thresholds to ensure safety.

3.2 Spectral Risk Measures

In order to reduce the likelihood of encountering the worst-case scenario, it is required to incorporate risk measures into constraints. Conditional value at risk (CVaR) is a widely used risk measure in finance [20], and it has been further developed into a spectral risk measure in [1], which provides a more comprehensive framework. A spectral risk measure with spectrum σ is defined as:

$$\mathcal{R}_\sigma(X) := \int_0^1 F_X^{-1}(u)\sigma(u)du,$$

where σ is an increasing function, $\sigma \geq 0$, $\int_0^1 \sigma(u)du = 1$, and F_X is the cumulative density function (CDF) of the random variable X. By appropriately defining the function σ, various risk measures can be established, and examples are as follows:

$$\text{CVaR}_\alpha : \sigma(u) = \mathbf{1}_{u\geq\alpha}/(1-\alpha), \quad \text{Pow}_\alpha : \sigma(u) = u^{\alpha/(1-\alpha)}/(1-\alpha), \tag{1}$$

where $\alpha \in [0, 1)$ represents a risk level, and the measures become risk neutral when $\alpha = 0$. Furthermore, it is known that the spectral risk measure has the following dual form [7]:

$$\mathcal{R}_\sigma(X) = \inf_g \mathbb{E}[g(X)] + \int_0^1 g^*(\sigma(u))du, \tag{2}$$

where $g : \mathbb{R} \mapsto \mathbb{R}$ is an increasing convex function, and $g^*(y) := \sup_x xy - g(x)$ is the convex conjugate function of g. For example, CVaR is expressed as $\text{CVaR}_\alpha(X) = \inf_\beta \mathbb{E}[\frac{1}{\alpha}(X-\beta)_+] + \beta$, where $g(x) = (x-\beta)_+/(1-\alpha)$, and the integral part of the conjugate function becomes β. We define the following sub-risk measure corresponding to the function g as follows:

$$\mathcal{R}^g_\sigma(X) := \mathbb{E}[g(X)] + \int_0^1 g^*(\sigma(u))du. \tag{3}$$

Then, $\mathcal{R}_\sigma(X) = \inf_g \mathcal{R}^g_\sigma(X)$. Note that the integral part is independent of X and only serves as a constant value of risk. As a result, the sub-risk measure can be expressed using the expectation, so it provides computational advantages over the original risk measure in many operations. We will utilize this advantage to show optimality of the proposed method.

3.3 Augmented State Spaces

As mentioned by Zhang et al. [32], a risk-constrained RL problem is non-Markovian, so a history-conditioned policy is required to solve the problem. Also, Bastani et al. [6] showed that an optimal history-conditioned policy can be expressed as a Markov policy defined in a newly augmented state space, which incorporates the discounted sum of costs as part of the state. To follow these results, we define an augmented state as follows:

$$\bar{s}_t := (s_t, \{e_{i,t}\}_{i=1}^N, b_t), \text{ where } e_{i,0} = 0,\ e_{i,t+1} = (C_i(s_t, a_t, s_{t+1}) + e_{i,t})/\gamma,\ b_t = \gamma^t. \tag{4}$$

Then, the CMDP is modified as follows:

$$\bar{s}_{t+1} = (s_{t+1}, \{(c_{i,t} + e_{i,t})/\gamma\}_{i=1}^N, \gamma b_t), \text{ where } s_{t+1} \sim P(\cdot|s_t, a_t),\ c_{i,t} = C_i(s_t, a_t, s_{t+1}),$$

and the policy $\pi(\cdot|\bar{s})$ is defined on the augmented state space.

4 Proposed Method

The risk measure-constrained RL (RCRL) problem is defined as follows:

$$\max_\pi \mathbb{E}[G^\pi_R] \ \ \mathbf{s.t.}\ \mathcal{R}_{\sigma_i}(G^\pi_{C_i}) \leq d_i\ \forall i \in \{1, 2, ..., N\}, \tag{5}$$

where d_i is the threshold of the ith constraint. Due to the nonlinearity of the risk measure, achieving an optimal policy through policy gradient techniques can be challenging. To address this issue, we

Algorithm 1 Overview of Spectral-Risk-Constrained Policy Optimization (SRCPO)

Input: Spectrum function σ_i for $i \in \{1, ..., N\}$.
Parameterize g_i of σ_i defined in (3) using $\beta_i \in B \subset \mathbb{R}^{M-1}$. // Parameterization, Section 6.1.
for $t = 1$ **to** T **do**
  for $\boldsymbol{\beta} = \{\beta_1, ..., \beta_N\} \in B^N$ **do**
    Update $\pi_{\boldsymbol{\beta},t}$ using the proposed policy gradient method. // Inner problem, Section 5.2.
  end for
  Calculate the target distribution of ξ using $\pi_{\boldsymbol{\beta},t}$, and update ξ_t. // Outer problem, Section 6.2.
end for
Output: $\pi_{\boldsymbol{\beta},T}$, where $\boldsymbol{\beta} \sim \xi_T$.

propose solving the RCRL problem by decomposing it into two separate problems. Using a feasibility function $\mathcal{F}$,[2] the RCRL problem can be rewritten as follows:

$$\max_{\pi} \mathbb{E}[G_R^\pi] - \sum_{i=1}^N \mathcal{F}(\mathcal{R}_{\sigma_i}(G_{C_i}^\pi) - d_i) = \max_{\pi} \mathbb{E}[G_R^\pi] - \sum_{i=1}^N \inf_{g_i} \mathcal{F}(\mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^\pi) - d_i)$$
$$= \underbrace{\sup_{\{g_i\}_{i=1}^N} \underbrace{\max_{\pi} \mathbb{E}[G_R^\pi] - \sum_{i=1}^N \mathcal{F}(\mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^\pi) - d_i)}_{\textbf{(a)}},}_{\textbf{(b)}} \quad (6)$$

where **(a)** is the inner problem, **(b)** is the outer problem, and $\{g_i\}_{i=1}^N$ is denoted as a dual variable. It is known that if the cost functions are bounded and there exists a policy that satisfies the constraints of (5), there exists an optimal dual variable [7], which transforms the supremum into the maximum in (6). Then, the RCRL problem can be solved by **1)** finding optimal policies for various dual variables and **2)** selecting the dual variable corresponding to the policy that achieves the maximum expected reward return while satisfying the constraints.

The inner problem is then a safe RL problem with sub-risk measure constraints. However, although the sub-risk measure is expressed using the expectation as in (3), the function g causes nonlinearity, which makes difficult to develop a policy gradient method. To address this issue, we propose novel risk value functions that show linearity in the performance difference between two policies. Through these risk value functions, we propose a policy update rule with convergence guarantees for the inner problem in Section 5.2.

The search space of the outer problem is a function space, rendering it impractical. To address this challenge, we appropriately parameterize the function g_i using a parameter $\beta_i \in B \subset \mathbb{R}^{M-1}$ in Section 6.1. Then, we can solve the outer problem through a brute-force approach by searching the parameter space. However, it is computationally prohibitive, requiring exponential time regarding to the number of constraints and the dimension of the parameter space. Consequently, it is necessary to develop an efficient method for solving the outer problem. Given the difficulty of directly identifying an optimal β^*, we instead model a distribution ξ to estimate the probability of a given β being optimal. We then propose a method that ensures convergence to an optimal distribution ξ^*, which deterministically samples an optimal β^*, in Section 6.2.

By integrating the proposed methods to solve the inner and outer problems, we finally introduce an RCRL algorithm called *spectral-risk-constrained policy optimization (SRCPO)*. We can obtain an optimal β^* by solving the outer problem and a set of optimal policies $\{\pi_\beta^* \mid \beta \in B\}$ by solving the inner problem, ultimately achieving the optimal policy $\pi_{\beta^*}^*$ of (5). An overview of SRCPO is presented in Algorithm 1, and a detailed pseudo-code of the proposed method is described in Algorithm 2. In the subsequent sections, we will provide details of the proposed methods to solve both the inner and the outer problems, respectively.

5 Inner Problem: Safe RL with Sub-Risk Measure

In this section, we propose a policy gradient method to solve the inner problem, defined as:

$$\max_{\pi} \mathbb{E}[G_R^\pi] \;\; \textbf{s.t.}\; \mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^\pi) \leq d_i \;\forall i, \quad (7)$$

[2] $\mathcal{F}(x) := 0$ if $x \leq 0$ else ∞.

where constraints consist of the sub-risk measures, defined in (3). In the remainder of this section, we first introduce risk value functions to calculate the policy gradient of the sub-risk measures, propose a policy update rule, and finally demonstrate its convergence to an optimum in tabular settings.

5.1 Risk Value Functions

In order to derive the policy gradient of the sub-risk measure, we first define risk value functions. To this end, using the fact that $e_{i,t} = \sum_{j=0}^{t-1} \gamma^j c_{i,j}/\gamma^t$ and $b_t = \gamma^t$, where $\{e_{i,t}\}_{i=1}^N$ and b_t is defined in the augmented state $\bar{s}_t$, we expand the sub-risk measure of the ith cost as follows:

$$\mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^\pi) - \int_0^1 g_i{}^*(\sigma_i(u))du = \int_{-\infty}^{\infty} g_i(z) f_{G_{C_i}^\pi}(z)dz = \mathbb{E}_{\bar{s}_0\sim\rho}\left[\int_{-\infty}^{\infty} g_i(z) f_{Y_{C_i}^\pi(\bar{s}_0)}(z)dz\right]$$

$$= \mathop{\mathbb{E}}_{\rho,\pi,P}\left[\int_{-\infty}^{\infty} g_i(z) f_{\sum_{j=0}^{t-1}\gamma^j c_{i,j}+\gamma^t Y_{C_i}^\pi(\bar{s}_t)}(z)dz\right] = \mathop{\mathbb{E}}_{\rho,\pi,P}\left[\int_{-\infty}^{\infty} g_i(z) f_{b_t(e_{i,t}+Y_{C_i}^\pi(\bar{s}_t))}(z)dz\right] \ \forall t,$$

where f_X is the probability density function (pdf) of a random variable X. To capture the value of the sub-risk measure, we can define risk value functions as follows:

$$V_{i,g}^\pi(\bar{s}) := \int_{-\infty}^{\infty} g(z) f_{b(e_i+Y_{C_i}^\pi(\bar{s}))}(z)dz, \ Q_{i,g}^\pi(\bar{s},a) := \int_{-\infty}^{\infty} g(z) f_{b(e_i+Z_{C_i}^\pi(\bar{s},a))}(z)dz.$$

Using the convexity of g, we present the following lemma on the bounds of the risk value functions:

Lemma 5.1. *Assuming that a function g is differentiable, $V_{i,g}^\pi(\bar{s})$ and $Q_{i,g}^\pi(\bar{s},a)$ are bounded by $[g(e_i b), g(e_i b + bC_{\max}/(1-\gamma))]$, and $|Q_{i,g}^\pi(\bar{s},a) - V_{i,g}^\pi(\bar{s})| \le bC$, where $C = \frac{C_{\max}}{1-\gamma} g'(\frac{C_{\max}}{1-\gamma})$ and $\bar{s} = (s, \{e_i\}_{i=1}^N, b)$.*

The proof is provided in Appendix A. Lemma 5.1 means that the value of $Q_{i,g}^\pi(\bar{s},a) - V_{i,g}^\pi(\bar{s})$ is scaled by b. To eliminate this scale, we define a risk advantage function as follows:

$$A_{i,g}^\pi(\bar{s},a) := (Q_{i,g}^\pi(\bar{s},a) - V_{i,g}^\pi(\bar{s}))/b.$$

Now, we introduce a theorem on the difference in the sub-risk measure between two policies.

Theorem 5.2. *Given two policies, π and π', the difference in the sub-risk measure is:*

$$\mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^{\pi'}) - \mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^\pi) = \mathbb{E}_{d_\rho^{\pi'},\pi'}\left[A_{i,g_i}^\pi(\bar{s},a)\right]/(1-\gamma). \tag{8}$$

The proof is provided in Appendix A. Since Theorem 5.2 also holds for the optimal policy, it is beneficial for showing optimality, as done in policy gradient algorithms of traditional RL [4]. Additionally, to calculate the advantage function, we use a distributional critic instead of directly estimating the risk value functions. This process is described in detail in Section 7.

5.2 Policy Update Rule

In this section, we derive the policy gradient of the sub-risk measure and introduce a policy update rule to solve the inner problem. Before that, we first parameterize the policy as π_θ, where $\theta \in \Theta$, and denote the objective function as $\mathbb{E}[G_R^{\pi_\theta}] = J_R(\pi) = J_R(\theta)$ and the ith constraint function as $\mathcal{R}_{\sigma_i}^{g_i}(G_{C_i}^{\pi_\theta}) = J_{C_i}(\pi) = J_{C_i}(\theta)$. Then, using (8) and the fact that $\mathbb{E}_{a\sim\pi(\cdot|\bar{s})}[A_{i,g}^\pi(\bar{s},a)] = 0$, the policy gradient of the ith constraint function is derived as follows:

$$\nabla_\theta J_{C_i}(\theta) = \mathbb{E}_{d_\rho^{\pi_\theta},\pi_\theta}[A_{i,g_i}^{\pi_\theta}(\bar{s},a)\nabla_\theta \log(\pi_\theta(a|\bar{s}))]/(1-\gamma). \tag{9}$$

Now, we propose a policy update rule as follows:

$$\theta_{t+1} = \begin{cases} \theta_t + \zeta_t F^\dagger(\theta_t)\nabla_\theta \left(J_R(\theta) - \zeta_t \sum_{i=1}^N \lambda_{t,i} J_{C_i}(\theta)\right)\big|_{\theta=\theta_t} & \text{if constraints are satisfied,} \\ \theta_t + \zeta_t F^\dagger(\theta_t)\nabla_\theta \left(\zeta_t \nu_t J_R(\theta) - \sum_{i=1}^N \lambda_{t,i} J_{C_i}(\theta)\right)\big|_{\theta=\theta_t} & \text{otherwise,} \end{cases}$$

where ζ_t is a learning rate, $F(\theta)$ is Fisher information matrix, $A^\dagger$ is Moore-Penrose inverse matrix of A, $\{\lambda_{t,1}, ..., \lambda_{t,N}, \nu_t\} \subset [0, \lambda_{\max}]$ are weights, and $\lambda_{t,i} = 0$ for $i \in \{i | J_{C_i}(\theta) \le d_i\}$ and $\sum_i \lambda_{t,i} = 1$ when the constraints are not satisfied. There are various possible strategies for determining $\lambda_{t,i}$ and ν_t, which makes the proposed method generalize many primal approach-based safe RL algorithms [2, 15, 25]. Several options for determining $\lambda_{t,i}$ and ν_t, as well as our strategy, are described in Appendix B.

5.3 Convergence Analysis

We analyze the convergence of the proposed policy update rule in a tabular setting where both the augmented state space and action space are finite. To demonstrate convergence in this setting, we use the softmax policy parameterization [4] as follows:

$$\pi_\theta(a|\bar{s}) := \exp\theta(\bar{s},a)/\left(\sum_{a'}\exp\theta(\bar{s},a')\right)\ \forall(\bar{s},a)\in\bar{S}\times A.$$

Then, we can show that the policy converges to an optimal policy if updated by the proposed method.

Theorem 5.3. *Assume that there exists a feasible policy π_f such that $J_{C_i}(\pi_f) \le d_i - \eta$ for $\forall i$, where $\eta > 0$. Then, if a policy is updated by the proposed update rule with a learning rate ζ_t that follows the Robbins-Monro condition [19], it will converge to an optimal policy of the inner problem.*

The proof is provided in Appendix A.1, and the assumption of the existence of a feasible policy is commonly used in convergence analysis in safe RL [5, 12, 15]. Through this result, we can also demonstrate convergence of existing primal approach-based safe RL methods, such as CPO [2], PCPO [27], and P3O [31],[3] by identifying proper $\lambda_{t,i}$ and ν_t strategies for these methods. We also provide an analysis of the convergence rate of the proposed method in Appendix A.2, employing a similar approach used in [25].

6 Outer Problem: Dual Optimization for Spectral Risk Measure

In this section, we propose a method for solving the outer problem that can be performed concurrently with the policy update of the inner problem. To this end, since the domain of the outer problem is a function space, we first introduce a procedure to parameterize the functions, allowing the problem to be practically solved. Subsequently, we propose a method to find an optimal solution for the parameterized outer problem.

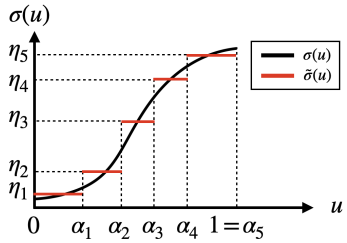


Figure 1: Discretization of spectrum.

6.1 Parameterization

Searching the function space of g is challenging. To address this issue, we discretize the spectrum function as illustrated in Figure 1 since the discretization allows g to be expressed using a finite number of parameters. The discretized spectrum is then formulated as:

$$\sigma(u) \approx \tilde{\sigma}(u) := \eta_1 + \sum_{j=1}^{M-1}(\eta_{j+1}-\eta_j)\mathbf{1}_{\mathrm{u}\ge\alpha_\mathrm{j}}, \tag{10}$$

where $0 \le \eta_j \le \eta_{j+1}$, $0 \le \alpha_j \le \alpha_{j+1} \le 1$, and M is the number of discretizations. The proper values of η_j and α_j can be achieved by minimizing the norm of the function space, expressed as:

$$\min_{\{\eta_j\}_{j=1}^{M},\{\alpha_j\}_{j=1}^{M-1}} \int_0^1 |\sigma(u)-\tilde{\sigma}(u)|du \quad \mathbf{s.t.}\ \int_0^1 \tilde{\sigma}(u)du = 1. \tag{11}$$

Now, we can show that the difference between the original risk measure and its discretized version is bounded by the inverse of the number of discretizations.

Lemma 6.1 (Approximation Error)**.** *The difference between the original risk measure and its discretized version is:*

$$|\mathcal{R}_\sigma(X) - \mathcal{R}_{\tilde{\sigma}}(X)| \le C_{\max}\sigma(1)/((1-\gamma)M).$$

The proof is provided in Appendix A. Lemma 6.1 means that as the number of discretizations increases, the risk measure converges to the original one. Note that the discretized version of the risk measure is also a spectral risk measure, indicating that the discretization process projects the original risk measure into a specific class of spectral risk measures. Now, we introduce a parameterization method as detailed in the following theorem:

[3] Note that these safe RL algorithms are designed to handle expectation-based constraints, so they are not suitable for RCRL.

Theorem 6.2. *Let us parameterize the function g using a parameter $\beta \in B \subset \mathbb{R}^{M-1}$ as:*

$$g_\beta(x) := \eta_1 x + \sum_{j=1}^{M-1} (\eta_{j+1} - \eta_j)(x - \beta[j])_+,$$

where $\beta[j] \leq \beta[j+1]$ for $j \in \{1, .., M-2\}$. Then, the following is satisfied:

$$\mathcal{R}_{\tilde{\sigma}}(X) = \inf_\beta \mathcal{R}^\beta_{\tilde{\sigma}}(X), \text{ where } \mathcal{R}^\beta_{\tilde{\sigma}}(X) := \mathbb{E}[g_\beta(X)] + \int_0^1 g^*_\beta(\tilde{\sigma}(u))du.$$

According to Remark 2.7 in [7], the infimum function in (2) exists and can be expressed as an integral of the spectrum function if the reward and cost functions are bounded. By using this fact, Theorem 6.2 can be proved, and details are provided in Appendix A. Through this parameterization, the RCRL problem is now expressed as follows:

$$\sup_{\{\beta_i\}_{i=1}^N} \max_\pi \mathbb{E}[G^\pi_R] - \sum_{i=1}^N \mathcal{F}(\mathcal{R}^{\beta_i}_{\tilde{\sigma}_i}(G^\pi_{C_i}) - d_i), \tag{12}$$

where $\beta_i \in B \subset \mathbb{R}^{M-1}$ is a parameter of the ith constraint. In the remainder of the paper, we denote the constraint function $\mathcal{R}^{\beta_i}_{\tilde{\sigma}_i}(G^\pi_{C_i})$ as $J_{C_i}(\pi; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^N$, and the policy for $\boldsymbol{\beta}$ as $\pi_{\boldsymbol{\beta}}$.

6.2 Optimization

To obtain the supremum of $\boldsymbol{\beta}$ in (12), a brute-force search can be used, but it demands exponential time relative to the dimension of $\boldsymbol{\beta}$. Alternatively, $\boldsymbol{\beta}$ can be directly optimized via gradient descent; however, due to the difficulty in confirming the convexity of the outer problem, optimal convergence cannot be assured. To resolve this issue, we instead propose to find a distribution on $\boldsymbol{\beta}$, called a *sampler* $\xi \in \mathcal{P}(B)^N$, that outputs the likelihood that a given $\boldsymbol{\beta}$ is optimal.

For detailed descriptions of the sampler ξ, the probability of sampling $\boldsymbol{\beta}$ is expressed as $\xi(\boldsymbol{\beta}) = \prod_{i=1}^N \xi[i](\beta_i)$, where $\xi[i]$ is the ith element of ξ, and β_i is $\boldsymbol{\beta}[i]$. The sampling process is denoted by $\boldsymbol{\beta} \sim \xi$, where each component is sampled according to $\beta_i \sim \xi[i]$. Implementation of this sampling process is similar to the stick-breaking process [23] due to the condition $\beta_i[j] \leq \beta_i[j+1]$ for $j \in \{1, ..., M-2\}$. Initially, $\beta_i[1]$ is sampled within $[0, C_{\max}/(1-\gamma)]$, and subsequent values $\beta_i[j+1]$ are sampled within $[\beta_i[j], C_{\max}/(1-\gamma)]$. Then, our target is to find the following optimal distribution:

$$\xi^*(\boldsymbol{\beta}) \begin{cases} \geq 0 & \text{if } \boldsymbol{\beta} \in \{\boldsymbol{\beta} | J_R(\pi^*_{\boldsymbol{\beta}}) = J_R(\pi^*_{\boldsymbol{\beta}^*})\}, \\ = 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\boldsymbol{\beta}^*$ is an optimal solution of the outer problem (12). Once we obtain the optimal distribution, $\boldsymbol{\beta}^*$ can be achieved by sampling from ξ^*.

In order to obtain the optimal distribution, we propose a novel update process by parameterizing the sampler and building a loss function. First, we parameterize the sampler as ξ_ϕ, where $\phi \in \Phi$, and define the following function:

$$J(\pi; \boldsymbol{\beta}) := J_R(\pi) - K \sum_i (J_{C_i}(\pi; \boldsymbol{\beta}) - d_i)_+.$$

It is known that for sufficiently large $K > 0$, the optimal policy of the inner problem, $\pi^*_{\boldsymbol{\beta}}$, is also the solution of $\max_\pi J(\pi; \boldsymbol{\beta})$, which means $J_R(\pi^*_{\boldsymbol{\beta}}) = \max_\pi J(\pi; \boldsymbol{\beta})$ [31]. Using this fact, we build a target distribution for the sampler as: $\bar{\xi}_t(\boldsymbol{\beta}) \propto \exp(J(\pi_{\boldsymbol{\beta},t}; \boldsymbol{\beta}))$, where $\pi_{\boldsymbol{\beta},t}$ is the policy for $\boldsymbol{\beta}$ at time-step t. Then, we define a loss function using the cross-entropy (H) as follows:

$$\mathcal{L}_t(\phi) := H(\xi_\phi, \bar{\xi}_t) = -\mathbb{E}_{\boldsymbol{\beta} \sim \xi_\phi}[\log \bar{\xi}_t(\boldsymbol{\beta})] = -\mathbb{E}_{\boldsymbol{\beta} \sim \xi_\phi}[J(\pi_{\boldsymbol{\beta},t}; \boldsymbol{\beta})],$$
$$\nabla_\phi \mathcal{L}_t(\phi) = -\mathbb{E}_{\boldsymbol{\beta} \sim \xi_\phi}[\nabla_\phi \log(\xi_\phi(\boldsymbol{\beta})) J(\pi_{\boldsymbol{\beta},t}; \boldsymbol{\beta})].$$

Finally, we present an update rule along with its convergence property in the following theorem.

Theorem 6.3. *Let us assume that the space of $\boldsymbol{\beta}$ is finite and update the sampler according to the following equation:*

$$\phi_{t+1} = \phi_t - \zeta F^\dagger(\phi_t) \nabla_\phi \mathcal{L}_t(\phi)|_{\phi=\phi_t}, \tag{14}$$

where ζ is a learning rate, and $F^\dagger(\phi)$ is the pseudo-inverse of Fisher information matrix of ξ_ϕ. Then, under a softmax parameterization, the sampler converges to an optimal distribution defined in (13).

The proof is provided in Appendix A. As the loss function consists of the current policy $\pi_{\boldsymbol{\beta},t}$, the sampler can be updated simultaneously with the policy update.

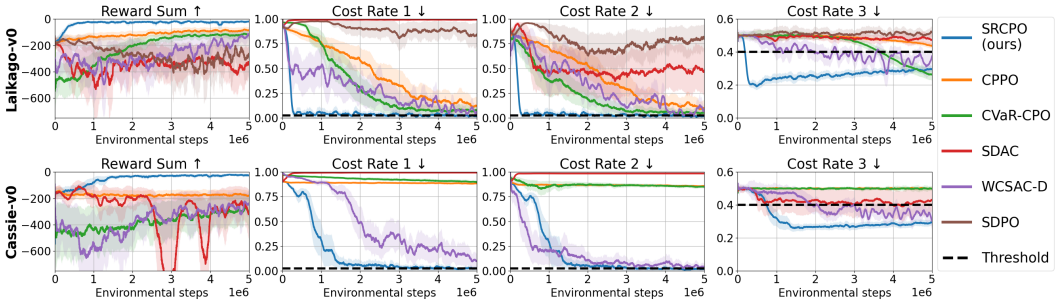


Figure 2: **Training curves of the legged robot locomotion tasks.** The upper graph shows results for the quadrupedal robot, and the lower one is for the bipedal robot. The solid line in each graph represents the average of each metric, and the shaded area indicates the standard deviation scaled by 0.5. The results are obtained by training each algorithm with five random seeds.

7 Practical Implementation

In order to calculate the policy gradient defined in (9), it is required to estimate the risk value functions. Instead of modeling them directly, we use quantile distributional critics [11] $Z_{C_i,\psi}(\bar{s}, a; \boldsymbol{\beta}) : \bar{S} \times A \times B^N \mapsto \mathbb{R}^L$, which approximate the pdf of $Z_{C_i}^{\pi_{\boldsymbol{\beta}}}(\bar{s}, a)$ using L Dirac delta functions and are parameterized by $\psi \in \Psi$. Then, the risk value function can be approximated as:

$$Q_{i,g}^{\pi_{\boldsymbol{\beta}}}(\bar{s}, a) \approx \sum_{l=1}^{L} g(be_i + bZ_{C_i,\psi}(\bar{s}, a; \boldsymbol{\beta})[l])/L,$$

and $V_{i,g}^{\pi_{\boldsymbol{\beta}}}(\bar{s})$ can be achieved from $\mathbb{E}_{a\sim\pi_{\boldsymbol{\beta}}(\cdot|\bar{s})}[Q_{i,g}^{\pi_{\boldsymbol{\beta}}}(\bar{s}, a)]$. The distributional critics are trained to minimize the quantile regression loss [11] and details are referred to Appendix C. Additionally, to implement ξ_ϕ, we use a truncated normal distribution $\mathcal{N}(\mu, \sigma^2, a, b)$ [9], where $[a, b]$ is the sampling range. Then, the sampling process of ξ_ϕ can be implemented as follows:

$$\beta_i[j] = \sum_{k=1}^{j} \Delta\beta_i[k], \text{ where } \Delta\beta_i[j] \sim \mathcal{N}(\mu_{i,\phi}[j], \sigma_{i,\phi}^2[j], 0, C_{\max}/(1-\gamma)) \ \forall i \text{ and } \forall j.$$

Additional details of the practical implementation are provided in Appendix C.

8 Experiments and Results

Tasks. The experiments are conducted on the Safety Gymnasium tasks [14] with a single constraint and the legged robot locomotion tasks [15] with multiple constraints. In the Safety Gymnasium, two robots—point and car—are used to perform two tasks: a goal task, which involves controlling the robot to reach a target location, and a button task, which involves controlling the robot to press a designated button. In these tasks, a cost is incurred when the robot collides with an obstacle. In the legged robot locomotion tasks, bipedal and quadrupedal robots are controlled to track a target velocity while satisfying three constraints related to body balance, body height, and foot contact timing. For more details, please refer to Appendix D.

Baselines. Many RCRL algorithms use CVaR to define risk constraints [15, 29, 32]. Accordingly, the proposed method is evaluated and compared with other algorithms under the CVaR constraints with $\alpha = 0.75$. Experiments on other risk measures are performed in Section 8.1. The baseline algorithms are categorized into three types based on their approach to estimating risk measures. First, CVaR-CPO [32] and CPPO [29] utilize auxiliary variables to estimate CVaR. Second, WCSAC-Dist [26] and SDPO [30] approximate the risk measure $\mathcal{R}_\sigma(G_C^\pi)$ with an expected value $\mathbb{E}_{s\sim\mathcal{D}}[\mathcal{R}_\sigma(Y_C^\pi(s))]$. Finally, SDAC [15] approximates G_C^π as a Gaussian distribution and uses the mean and standard deviation of G_C^π to estimate CVaR. The hyperparameters and network structure of each algorithm are detailed in Appendix D. Note that both CVaR-CPO and the proposed method, SRCPO, use the augmented state space. However, for a fair comparison with other algorithms, the policy and critic networks of these methods are modified to operate on the original state space.

Results. Figures 2 and 3 show the training curves for the locomotion tasks and the Safety Gymnasium tasks, respectively. The reward sum in the figures refers to the sum of rewards within an episode,

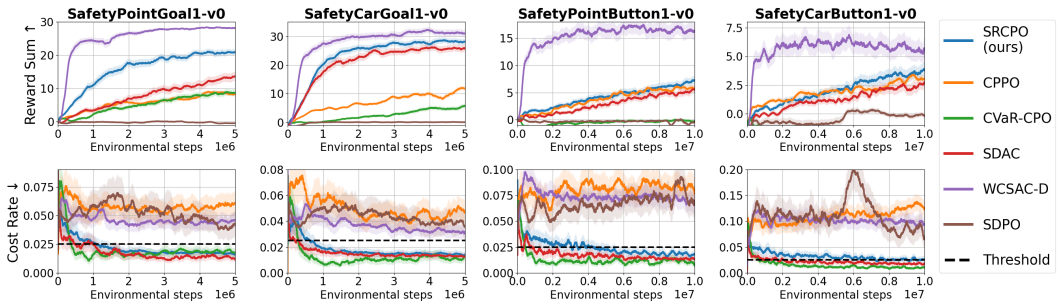


Figure 3: **Training curves of the Safety Gymnasium tasks.** The results for each task are displayed in columns, titled with the task name. The solid line represents the average of each metric, and the shaded area indicates the standard deviation scaled by 0.2. The results are obtained by training each algorithm with five random seeds.

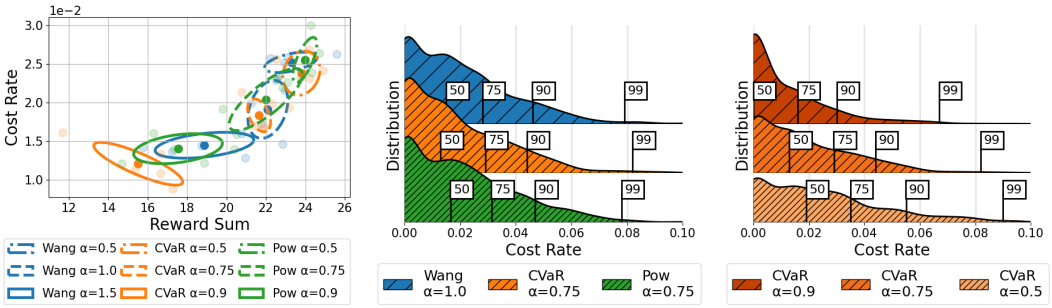


Figure 4: **(Left)** A correlation graph between cost rate and reward sum for policies trained in the point goal task under various risk measure constraints. The results are achieved by training policies with five random seeds for each risk measure and risk level. The center and radius of each ellipse show the average and standard deviation of the results from the five seeds, respectively. **(Middle)** Distribution graphs of the cost rate under different risk measure constraints. Locations of several percentiles (from the 50th to the 99th) are marked on the plot. The risk level of each risk measure is selected to have a similar cost rate. After training a policy in the point goal task, cost distributions have been collected by rolling out the trained policy across 500 episodes. **(Right)** Distribution graphs of the cost rate with different risk levels, α, under the CVaR constraint.

while the cost rate is calculated as the sum of costs divided by the episode length. In all tasks, the proposed method achieves the highest rewards among methods whose cost rates are below the specified thresholds. These results are likely because only the proposed method guarantees optimality. Specifically in the locomotion tasks, an initial policy often struggles to stabilize the balance of the robot, resulting in high costs from the start. Given this challenge, it is more susceptible to falling into local optima compared to other tasks, which enables the proposed method to outperform other baseline methods. Note that WCSAC-Dist shows the highest rewards in the Safety Gymnasium tasks, but the cost rates exceed the specified thresholds. This issue seems to arise from the approach to estimating risk constraints. WCSAC-Dist estimates the constraints based on the expected risk for each state, but it is lower than the original risk measure, leading to constraint violations.

8.1 Study on Various Risk Measures

In this section, we analyze the results when constraints are defined using various risk measures. To this end, we train policies in the point goal task under constraints based on the CVaR_α and Pow_α risk measures defined in (2), as well as the Wang risk measure [24]. Although the Wang risk measure is not a spectral but a distortion risk measure, our parameterization method introduced in Section 6.1 enables it to be approximated as a spectral risk measure, and the visualization of this process is provided in Appendix E. We conduct experiments with three risk levels for each risk measure and set the constraint threshold as 0.025. Evaluation results are presented in Figure 4, and training curves are provided in Appendix E. Figure 4 (Right) shows intuitive results that increasing the risk level effectively reduces the likelihood of incurring high costs. Similarly, Figure 4 (Left) presents the trend across all risk measures, indicating that higher risk levels correspond to lower cost rates and decreased reward performance. Finally, Figure 4 (Middle) exhibits the differences in how each risk measure addresses worst-case scenarios. In the spectrum formulation defined in (1), CVaR applies a

uniform penalty to the tail of the cost distribution above a specified percentile, whereas measures like Wang and Pow impose a heavier penalty at higher percentiles. As a result, because CVaR imposes a relatively milder penalty on the worst-case outcomes compared to other risk measures, it shows the largest intervals between the 50th and 99th percentiles.

9 Conclusions and Limitations

In this work, we introduced a spectral-risk-constrained RL algorithm that ensures convergence and optimality in a tabular setting. Specifically, in the inner problem, we proposed a generalized policy update rule that can facilitate the development of a new safe RL algorithm with convergence guarantees. For the outer problem, we introduced a notion called *sampler*, which enhances training efficiency by concurrently training with the inner problem. Through experiments in continuous control tasks, we empirically demonstrated the superior performance of the proposed method and its capability to handle various risk measures. However, convergence to an optimum is shown only in a tabular setting, so future research may focus on extending these results to linear MDPs or function approximation settings. Furthermore, since our approach can be applied to risk-sensitive RL, future work can also implement the proposed method in this area.

Acknowledgments and Disclosure of Funding

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01190, [SW Star Lab] Robot Learning: Efficient, Safe, and Socially-Acceptable Machine Learning, 70%) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00211357, Smart Assembler: Robot Active Learning for Unseen Parts Assembly, 30%)

References

[1] Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of International Conference on Machine Learning*, pages 22–31, 2017.

[3] Patrick Adjei, Norman Tasfi, Santiago Gomez-Rosero, and Miriam AM Capretz. Safe reinforcement learning for arm manipulation with constrained Markov decision process. *Robotics*, 13(4): 63, 2024.

[4] 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.

[5] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3682–3689, 2022.

[6] Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 36259–36269, 2022.

[7] Nicole Bäuerle and Alexander Glauner. Minimizing spectral risk measures applied to Markov decision processes. *Mathematical Methods of Operations Research*, 94(1):35–69, 2021.

[8] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of International Conference on Machine Learning*, pages 449–458, 2017.

[9] John Burkardt. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, 1(35):58, 2014.

[10] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18 (1):6070–6120, 2017.

[11] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[12] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 8378–8390, 2020.

[13] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of International Conference on Machine Learning*, pages 1861–1870, 2018.

[14] Jiaming Ji, Borong Zhang, Xuehai Pan, Jiayi Zhou, Juntao Dai, and Yaodong Yang. Safety-gymnasium. *GitHub repository*, 2023.

[15] Dohyeong Kim, Kyungjae Lee, and Songhwai Oh. Trust region-based safe distributional reinforcement learning for multiple constraints. In *Advances in Neural Information Processing Systems*, 2023.

[16] Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Scale-invariant gradient aggregation for constrained multi-objective reinforcement learning. *arXiv preprint arXiv:2403.00282*, 2024.

[17] Puze Liu, Haitham Bou-Ammar, Jan Peters, and Davide Tateo. Safe reinforcement learning on the constraint manifold: Theory and applications. *arXiv preprint arXiv:2404.09080*, 2024.

[18] Yongshuai Liu, Jiaxin Ding, and Xin Liu. IPO: Interior-point policy optimization under constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4940–4947, 2020.

[19] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[20] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

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

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

[23] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

[24] Shaun S Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, pages 15–36, 2000.

[25] Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *Proceedings of International Conference on Machine Learning*, pages 11480–11491, 2021.

[26] Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3):859–887, 2023.

[27] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. Projection-based constrained policy optimization. In *Proceedings of International Conference on Learning Representations*, 2020.

[28] Tsung-Yen Yang, Tingnan Zhang, Linda Luu, Sehoon Ha, Jie Tan, and Wenhao Yu. Safe reinforcement learning for legged locomotion. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2454–2461, 2022.

[29] Chengyang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2022.

[30] Jianyi Zhang and Paul Weng. Safe distributional reinforcement learning. In *Proceedings of Distributed Artificial Intelligence*, pages 107–128, 2022.

[31] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Penalized proximal policy optimization for safe reinforcement learning. *arXiv preprint arXiv:2205.11814*, 2022.

[32] Qiyuan Zhang, Shu Leng, Xiaoteng Ma, Qihan Liu, Xueqian Wang, Bin Liang, Yu Liu, and Jun Yang. CVaR-constrained policy optimization for safe reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

A Proofs

Lemma 5.1. *Assuming that a function g is differentiable, $V^{\pi}_{i,g}(\bar{s})$ and $Q^{\pi}_{i,g}(\bar{s},a)$ are bounded by $[g(e_i b), g(e_i b + bC_{\max}/(1-\gamma))]$, and $|Q^{\pi}_{i,g}(\bar{s},a) - V^{\pi}_{i,g}(\bar{s})| \leq bC$, where $C = \frac{C_{\max}}{1-\gamma} g'(\frac{C_{\max}}{1-\gamma})$ and $\bar{s} = (s, \{e_i\}_{i=1}^{N}, b)$.*

Proof. Since $0 \leq Y^{\pi}_{C_i}(\bar{s}), Z^{\pi}_{C_i}(\bar{s},a) \leq C_{\max}/(1-\gamma)$,

$$f_{be_i + bY^{\pi}_{C_i}(\bar{s})}(z) = f_{be_i + bZ^{\pi}_{C_i}(\bar{s},a)}(z) = 0,$$

when $z < be_i$ or $z > be_i + bC_{\max}/(1-\gamma)$. Using the fact that g is an increasing and convex function,

$$g(be_i) \leq V^{\pi}_{i,g}(\bar{s}) = \int_{be_i}^{be_i + b\frac{C_{\max}}{1-\gamma}} g(z) f_{be_i + bY^{\pi}_{C_i}(\bar{s})}(z) dz \leq g(be_i + bC_{\max}/(1-\gamma)).$$

Likewise, $Q^{\pi}_{i,g}(\bar{s},a)$ also satisfies the above property. As a result,

$$V^{\pi}_{i,g}(\bar{s}),\ Q^{\pi}_{i,g}(\bar{s},a) \in [g(be_i), g(be_i + bC_{\max}/(1-\gamma))].$$

Then, the maximum of $|Q^{\pi}_{i,g}(\bar{s},a) - V^{\pi}_{i,g}(\bar{s})|$ is $g(be_i + bC_{\max}/(1-\gamma)) - g(be_i)$. Since $b = \gamma^t$ and $e_i = \sum_{k=0}^{t-1} \gamma^k C_i(s_k, a_k, s_{k+1})/\gamma^t$, where t is the current time step,

$$be_i + bC_{\max}/(1-\gamma) \leq C_{\max}/(1-\gamma).$$

Consequently,

$$\begin{aligned} |Q^{\pi}_{i,g}(\bar{s},a) - V^{\pi}_{i,g}(\bar{s})| &\leq g(be_i + bC_{\max}/(1-\gamma)) - g(be_i) \\ &\leq b\frac{C_{\max}}{1-\gamma} g'\left(be_i + b\frac{C_{\max}}{1-\gamma}\right) \leq b\frac{C_{\max}}{1-\gamma} g'\left(\frac{C_{\max}}{1-\gamma}\right). \end{aligned}$$

□

Theorem 5.2. *Given two policies, π and π', the difference in the sub-risk measure is:*

$$\mathcal{R}^{g_i}_{\sigma_i}(G^{\pi'}_{C_i}) - \mathcal{R}^{g_i}_{\sigma_i}(G^{\pi}_{C_i}) = \mathbb{E}_{d^{\pi'}_{\rho},\pi'}\left[A^{\pi}_{i,g_i}(\bar{s},a)\right]/(1-\gamma). \tag{8}$$

Proof. Since the following equation is satisfied:

$$\begin{aligned} \int_{-\infty}^{\infty} g(z) f_{be_i + bZ^{\pi}_{C_i}(\bar{s},a)}(z) dz &= \mathbb{E}_{\bar{s}' \sim P(\cdot|\bar{s},a)}\left[\int_{-\infty}^{\infty} g(z) f_{b(e_i + C_i(s,a,s')) + \gamma b Y^{\pi}_{C_i}(\bar{s}')}(z) dz\right] \\ &= \mathbb{E}_{\bar{s}' \sim P(\cdot|\bar{s},a)}\left[\int_{-\infty}^{\infty} g(z) f_{b' e'_i + b' Y^{\pi}_{C_i}(\bar{s}')}(z) dz\right], \end{aligned}$$

we can say $Q^{\pi}_{i,g}(\bar{s},a) = \mathbb{E}_{\bar{s}' \sim P(\cdot|\bar{s},a)}[V^{\pi}_{i,g}(\bar{s}')]$. Using this property,

$$\begin{aligned} \mathbb{E}_{d^{\pi'}_{\rho},\pi'}\left[A^{\pi}_{i,g_i}(\bar{s},a)\right]/(1-\gamma) &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi}_{i,g_i}(\bar{s}_t, a_t)\right] \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} Q^{\pi}_{i,g_i}(\bar{s}_t, a_t) - V^{\pi}_{i,g_i}(\bar{s}_t)\right] \\ &= \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \left(V^{\pi}_{i,g_i}(\bar{s}_{t+1}) - V^{\pi}_{i,g_i}(\bar{s}_t)\right)\right] \\ &= \lim_{t\to\infty} \mathbb{E}_{\pi'}\left[V^{\pi}_{i,g_i}(\bar{s}_t)\right] - \mathbb{E}_{\bar{s}_0 \sim \rho}\left[V^{\pi}_{i,g_i}(\bar{s}_0)\right] \\ &= \lim_{t\to\infty} \mathbb{E}_{\pi'}\left[V^{\pi}_{i,g_i}(\bar{s}_t)\right] - \mathcal{R}^{g_i}_{\sigma_i}(G^{\pi}_{C_i}) + \int_0^1 {g_i}^*(\sigma_i(u)) du. \end{aligned} \tag{15}$$

Using the definition of $V^{\pi}_{i,g_i}(\bar{s}_t)$ and the dominated convergence theorem,

$$\begin{aligned}\lim_{t\to\infty}\mathbb{E}_{\pi'}\left[V^{\pi}_{i,g_i}(\bar{s}_t)\right] &= \lim_{t\to\infty}\mathbb{E}_{\pi'}\left[\int_{-\infty}^{\infty} g_i(z) f_{b_t e_{i,t}+b_t Y^{\pi}_{C_i}(\bar{s}_t)}(z)dz\right] \\ &= \lim_{t\to\infty}\mathbb{E}_{\pi'}\left[\int_{-\infty}^{\infty} g_i(z) f_{\sum_{k=0}^{t-1}\gamma^k C_i(s_k,a_k,s_{k+1})+\gamma^t Y^{\pi}_{C_i}(\bar{s}_t)}(z)dz\right] \\ &= \mathbb{E}_{\pi'}\left[\int_{-\infty}^{\infty} g_i(z) f_{G^{\pi'}_{C_i}}(z)dz\right] = \mathcal{R}^{g_i}_{\sigma_i}(G^{\pi'}_{C_i}) - \int_0^1 {g_i}^*(\sigma_i(u))du.\end{aligned} \tag{16}$$

By combining (15) and (16),

$$\begin{aligned}\mathbb{E}_{d^{\pi'}_{\rho},\pi'}\left[A^{\pi}_{i,g_i}(\bar{s},a)\right]/(1-\gamma) &= \lim_{t\to\infty}\mathbb{E}_{\pi'}\left[V^{\pi}_{i,g_i}(\bar{s}_t)\right] - \mathcal{R}^{g_i}_{\sigma_i}(G^{\pi}_{C_i}) + \int_0^1 {g_i}^*(\sigma_i(u))du \\ &= \mathcal{R}^{g_i}_{\sigma_i}(G^{\pi'}_{C_i}) - \mathcal{R}^{g_i}_{\sigma_i}(G^{\pi}_{C_i}).\end{aligned}$$

□

Lemma 6.1 (Approximation Error)**.** *The difference between the original risk measure and its discretized version is:*

$$|\mathcal{R}_{\sigma}(X) - \mathcal{R}_{\tilde{\sigma}}(X)| \le C_{\max}\sigma(1)/((1-\gamma)M).$$

Proof. The difference is:

$$|\mathcal{R}_{\sigma}(X) - \mathcal{R}_{\tilde{\sigma}}(X)| \le \int_0^1 |F_X^{-1}(u)||\sigma(u)-\tilde{\sigma}(u)|du \le ||\sigma-\tilde{\sigma}||_1||F_X^{-1}||_{\infty} = ||\sigma-\tilde{\sigma}||_1\frac{C_{\max}}{1-\gamma}.$$

The value of (11) is $||\sigma-\tilde{\sigma}||_1$ and is smaller than the value of the following problem:

$$\min_{\eta_j,\alpha_j}\int_0^1 |\sigma(u)-\tilde{\sigma}(u)|du \quad \text{s.t. } \int_0^1 \tilde{\sigma}(u)du = 1,\ \frac{\sigma(1)(j-1)}{M} \le \eta_j \le \frac{\sigma(1)j}{M}\ \forall j, \tag{17}$$

since the search space is smaller than the original one. As the value of (17) is smaller than $\max_u|\sigma(u)-\tilde{\sigma}(u)| \le \sigma(1)/M$, $||\sigma-\tilde{\sigma}||_1 \le \sigma(1)/M$. Consequently,

$$|\mathcal{R}_{\sigma}(X) - \mathcal{R}_{\tilde{\sigma}}(X)| \le ||\sigma-\tilde{\sigma}||_1\frac{C_{\max}}{1-\gamma} \le \frac{C_{\max}\sigma(1)}{(1-\gamma)M}.$$

□

Theorem 6.2. *Let us parameterize the function g using a parameter $\beta \in B \subset \mathbb{R}^{M-1}$ as:*

$$g_{\beta}(x) := \eta_1 x + \sum_{j=1}^{M-1}(\eta_{j+1}-\eta_j)(x-\beta[j])_+,$$

where $\beta[j] \le \beta[j+1]$ for $j \in \{1,..,M-2\}$. Then, the following is satisfied:

$$\mathcal{R}_{\tilde{\sigma}}(X) = \inf_{\beta}\mathcal{R}^{\beta}_{\tilde{\sigma}}(X),\ \textit{where}\ \mathcal{R}^{\beta}_{\tilde{\sigma}}(X) := \mathbb{E}[g_{\beta}(X)] + \int_0^1 g^*_{\beta}(\tilde{\sigma}(u))du.$$

Proof. Since $g_{\beta}(x)$ is an increasing convex function, $\mathcal{R}_{\tilde{\sigma}}(X) \le \mathcal{R}^{g_{\beta}}_{\tilde{\sigma}}(X) =: \mathcal{R}^{\beta}_{\tilde{\sigma}}(X)$. In addition, according to Remark 2.7 in [1], there exists $\tilde{g}$ such that $\tilde{g} = \arg\inf_g \mathcal{R}^{g}_{\tilde{\sigma}}(X)$, and its derivative is expressed as $\tilde{g}' = \tilde{\sigma}(F_X(x))$. Using this fact, we can formulate $\tilde{g}$ by integrating its derivative as follows:

$$\tilde{g}(x) = \eta_1 x + \sum_{j=1}^{M-1}(\eta_{j+1}-\eta_j)(x-F_X^{-1}(\alpha_j))_+ + C,$$

where C is a constant. Since for any constant C, $\mathcal{R}^{\tilde{g}}_{\tilde{\sigma}}(X)$ has the same value due to the integral part of $\tilde{g}^*$ in (3), we set C as zero. Also, we can express $\tilde{g}(x)$ as $g_{\tilde{\beta}}(x)$, where $\tilde{\beta}[j]$ is $F_X^{-1}(\alpha_j)$ due to the definition of g_{β}. As a result, the following is satisfied:

$$\begin{aligned}\mathcal{R}_{\tilde{\sigma}}(X) = \mathcal{R}^{\tilde{g}}_{\tilde{\sigma}}(X) = \mathcal{R}^{\tilde{\beta}}_{\tilde{\sigma}}(X) \le \mathcal{R}^{\beta}_{\tilde{\sigma}}(X). \\ \Rightarrow \mathcal{R}_{\tilde{\sigma}}(X) = \inf_{\beta}\mathcal{R}^{\beta}_{\tilde{\sigma}}(X).\end{aligned}$$

□

Theorem 6.3. *Let us assume that the space of $\boldsymbol{\beta}$ is finite and update the sampler according to the following equation:*

$$\phi_{t+1} = \phi_t - \zeta F^{\dagger}(\phi_t)\nabla_{\phi}\mathcal{L}_t(\phi)|_{\phi=\phi_t}, \tag{14}$$

where ζ is a learning rate, and $F^{\dagger}(\phi)$ is the pseudo-inverse of Fisher information matrix of ξ_{ϕ}. Then, under a softmax parameterization, the sampler converges to an optimal distribution defined in (13).

Proof. Since we assume that the space of $\boldsymbol{\beta}$ is finite, the sampler can be expressed using the softmax parameterization as follows:

$$\xi_{\phi}(\boldsymbol{\beta}) := \frac{\exp(\phi(\boldsymbol{\beta}))}{\sum_{\boldsymbol{\beta}'}\exp(\phi(\boldsymbol{\beta}'))}.$$

If we update the parameter of the sampler as described in (14), the following is satisfied as in the natural policy gradient:

$$\phi_{t+1}(\boldsymbol{\beta}) = \phi_t(\boldsymbol{\beta}) + \zeta J(\pi_{\boldsymbol{\beta},t};\boldsymbol{\beta}) = \sum_{k=1}^{t}\zeta J(\pi_{\boldsymbol{\beta},k};\boldsymbol{\beta}).$$

Furthermore, due to Theorem 5.3, if a policy $\pi_{\boldsymbol{\beta},t}$ is updated by the proposed method, the policy converges to an optimal policy $\pi^*_{\boldsymbol{\beta},t}$. Then, if $\boldsymbol{\beta}$ is not optimal,

$$\lim_{T\to\infty}\sum_{t=1}^{T}\zeta\left(J(\pi_{\boldsymbol{\beta},t};\boldsymbol{\beta}) - J(\pi_{\boldsymbol{\beta}^*,t};\boldsymbol{\beta}^*)\right) = -\infty,$$

since $\lim_{t\to\infty} J(\pi_{\boldsymbol{\beta},t};\boldsymbol{\beta}) - J(\pi_{\boldsymbol{\beta}^*,t};\boldsymbol{\beta}^*) = J(\pi^*_{\boldsymbol{\beta}};\boldsymbol{\beta}) - J(\pi^*_{\boldsymbol{\beta}^*};\boldsymbol{\beta}^*) < 0$. As a result,

$$\lim_{t\to\infty}\xi_{\phi_t}(\boldsymbol{\beta}) = \lim_{t\to\infty}\exp(\phi_t(\boldsymbol{\beta}) - \phi_t(\boldsymbol{\beta}^*))/\sum_{\boldsymbol{\beta}'}\exp(\phi_t(\boldsymbol{\beta}') - \phi_t(\boldsymbol{\beta}^*))$$

$$= \lim_{t\to\infty}\exp\left(\sum_{k=1}^{t-1}\zeta\left(J(\pi_{\boldsymbol{\beta},k};\boldsymbol{\beta}) - J(\pi_{\boldsymbol{\beta}^*,k};\boldsymbol{\beta}^*)\right)\right)\Big/\sum_{\boldsymbol{\beta}'}\exp(\phi_t(\boldsymbol{\beta}') - \phi_t(\boldsymbol{\beta}^*)) = 0.$$

Thus, ξ_{ϕ_t} converges to an optimal sampler. □

A.1 Proof of Theorem 5.3

Our derivation is based on the policy gradient works by Agarwal et al. [4] and Xu et al. [25]. Due to the softmax policy parameterization, the following is satisfied:

$$\pi_{t+1}(a|\bar{s}) = \pi_t(a|\bar{s})\frac{\exp(\theta_{t+1}(\bar{s},a) - \theta_t(\bar{s},a))}{Z_t(\bar{s})}, \tag{18}$$

where $Z_t(\bar{s}) := \sum_{a\in A}\pi_t(a|\bar{s})\exp(\theta_{t+1}(\bar{s},a) - \theta_t(\bar{s},a))$. By the natural policy gradient,

$$\theta_{t+1} = \begin{cases}\theta_t + \zeta_t(A^t_R - \zeta_t\sum_i \lambda_{t,i}A^t_{i,g_i})/(1-\gamma) & \text{if } J_{C_i}(\theta_t) \le d_i\ \forall i,\\ \theta_t + \zeta_t(\zeta_t\nu_t A^t_R - \sum_i \lambda_{t,i}A^t_{i,g_i})/(1-\gamma) & \text{otherwise.}\end{cases} \tag{19}$$

Before proving Theorem 5.3, we first present several useful lemmas.

Lemma A.1. $\log Z_t(\bar{s})$ *is non-negative.*

Proof. Due to the advantage functions in (19), $\sum_{a\in A}\pi_t(a|\bar{s})(\theta_{t+1}(\bar{s},a) - \theta_t(\bar{s},a)) = 0$. Using the fact that a logarithmic function is concave,

$$\log Z_t(\bar{s}) = \log\sum_{a\in A}\pi_t(a|\bar{s})\exp(\theta_{t+1}(\bar{s},a) - \theta_t(\bar{s},a)) \ge \sum_{a\in A}\pi_t(a|\bar{s})(\theta_{t+1}(\bar{s},a) - \theta_t(\bar{s},a)) = 0.$$

As a result, $\log Z_t(\bar{s}) \ge 0$. □

Lemma A.2. $D_{\mathrm{TV}}(\pi_{t+1}(\cdot|\bar{s}), \pi_t(\cdot|\bar{s})) \le \max_a|\theta_{t+1}(\bar{s},a) - \theta_t(\bar{s},a)| \le ||\theta_{t+1} - \theta_t||_{\infty}$.

Proof. Using the Pinsker's inequality,

$$
\begin{aligned}
&D_{\mathrm{TV}}(\pi_{t+1}(\cdot|\bar{s}), \pi_t(\cdot|\bar{s})) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\pi_{t+1}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s}))} \\
&= \sqrt{\frac{1}{2} \sum_a \pi_{t+1}(a|\bar{s}) \log(\pi_{t+1}(a|\bar{s})/\pi_t(a|\bar{s}))} \\
&= \sqrt{\frac{1}{2} (\sum_a \pi_{t+1}(a|\bar{s})(\theta_{t+1}(\bar{s}, a) - \theta_t(\bar{s}, a)) - \log Z_t(\bar{s}))} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{1}{2} \sum_a \pi_{t+1}(a|\bar{s})(\theta_{t+1}(\bar{s}, a) - \theta_t(\bar{s}, a))} \\
&\overset{\text{(ii)}}{=} \sqrt{\frac{1}{2} \sum_a (\pi_{t+1}(a|\bar{s}) - \pi_t(a|\bar{s}))(\theta_{t+1}(\bar{s}, a) - \theta_t(\bar{s}, a))} \\
&\leq \sqrt{\max_a |\theta_{t+1}(\bar{s}, a) - \theta_t(\bar{s}, a)| D_{\mathrm{TV}}(\pi_{t+1}(a|\bar{s}), \pi_t(a|\bar{s}))},
\end{aligned}
$$

where (i) follows Lemma A.1, and (ii) follows the fact that $\sum_{a \in A} \pi_t(a|\bar{s})(\theta_{t+1}(\bar{s}, a) - \theta_t(\bar{s}, a)) = 0$. As a result,

$$
D_{\mathrm{TV}}(\pi_{t+1}(\cdot|\bar{s}), \pi_t(\cdot|\bar{s})) \leq \max_a |\theta_{t+1}(\bar{s}, a) - \theta_t(\bar{s}, a)| \leq ||\theta_{t+1} - \theta_t||_\infty.
$$

□

Theorem 5.3. *Assume that there exists a feasible policy π_f such that $J_{C_i}(\pi_f) \leq d_i - \eta$ for $\forall i$, where $\eta > 0$. Then, if a policy is updated by the proposed update rule with a learning rate ζ_t that follows the Robbins-Monro condition [19], it will converge to an optimal policy of the inner problem.*

Proof. We divide the analysis into two cases: 1) when the constraints are satisfied and 2) when the constraints are violated. First, let us consider the case where the constraints are satisfied. In this case, the policy is updated as follows:

$$
\theta_{t+1} = \theta_t + \frac{\zeta_t}{1 - \gamma} \left(A_R^t - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t \right). \tag{20}
$$

Using Theorem 5.2,

$$
\begin{aligned}
&J_R(\theta_{t+1}) - J_R(\theta_t) + \zeta_t \sum_i \lambda_{t,i} (J_{C_i}(\theta_t) - J_{C_i}(\theta_{t+1})) \\
&= \frac{1}{1 - \gamma} \sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_a \pi_{t+1}(a|\bar{s}) (A_R^t(\bar{s}, a) - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s}, a)) \\
&= \frac{1}{1 - \gamma} \sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_a (\pi_{t+1}(a|\bar{s}) - \pi_t(a|\bar{s})) (A_R^t(\bar{s}, a) - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s}, a)) \\
&\overset{\text{(i)}}{\leq} \frac{2}{1 - \gamma} \sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \max_a \left| A_R^t(\bar{s}, a) - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s}, a) \right| D_{\mathrm{TV}}(\pi_{t+1}(\cdot|\bar{s}), \pi_t(\cdot|\bar{s})) \\
&\overset{\text{(ii)}}{\leq} \frac{2||\theta_{t+1} - \theta_t||_\infty}{1 - \gamma} \sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \max_a \left| A_R^t(\bar{s}, a) - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s}, a) \right| \\
&\leq \frac{2||\theta_{t+1} - \theta_t||_\infty}{1 - \gamma} \left|\left| A_R^t - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t \right|\right|_\infty = \frac{2\zeta_t}{(1 - \gamma)^2} \left|\left| A_R^t - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t \right|\right|_\infty^2 \\
&\leq 2\zeta_t (R_{\max} + \zeta_t N \lambda_{\max} C_{\max})^2 / (1 - \gamma)^4.
\end{aligned} \tag{21}
$$

where (i) follows Hölder's inequality, and (ii) follows Lemma A.2. Using (18) and (20),

$$
\begin{aligned}
&J_R(\theta_{t+1}) - J_R(\theta_t) + \zeta_t \sum_i \lambda_{t,i}(J_{C_i}(\theta_t) - J_{C_i}(\theta_{t+1})) \\
&= \frac{1}{1-\gamma} \sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_{a\in A} \pi_{t+1}(a|\bar{s})(A_R^t(\bar{s},a) - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)) \\
&= \frac{1}{\zeta_t} \sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_{a\in A} \pi_{t+1}(a|\bar{s}) \log(\pi_{t+1}(a|\bar{s}) Z_t(\bar{s})/\pi_t(a|\bar{s})) \\
&= \frac{1}{\zeta_t} \sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) (D_{\mathrm{KL}}(\pi_{t+1}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s})) + \log Z_t(\bar{s})) \\
&\geq \frac{1}{\zeta_t} \sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) \log Z_t(\bar{s}) \overset{\text{(i)}}{\geq} \frac{1-\gamma}{\zeta_t} \sum_{\bar{s}\in\bar{S}} \rho(\bar{s}) \log Z_t(\bar{s}),
\end{aligned}
\tag{22}
$$

where (i) is from the fact that $d_\rho^\pi(\bar{s}) = (1-\gamma)\sum_t \gamma^t \mathbb{P}(\bar{s}_t = \bar{s}) \geq (1-\gamma)\rho(\bar{s})$. By combining (21) and (22), the following is satisfied for any ρ:

$$
\frac{1-\gamma}{\zeta_t} \sum_{\bar{s}\in\bar{S}} \rho(\bar{s}) \log Z_t(\bar{s}) \leq \frac{2\zeta_t}{(1-\gamma)^4} (R_{\max} + \zeta_t N \lambda_{\max} C_{\max})^2. \tag{23}
$$

Then, the following is satisfied for any policy $\hat{\pi}$:

$$
\begin{aligned}
&J_R(\hat{\pi}) - J_R(\pi_t) + \zeta_t \sum_i \lambda_{t,i}(J_{C_i}(\pi_t) - J_{C_i}(\hat{\pi})) \\
&= \frac{1}{1-\gamma} \sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s}) \sum_{a\in A} \hat{\pi}(a|\bar{s})(A_R^t(\bar{s},a) - \zeta_t \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)) \\
&= \frac{1}{\zeta_t} \sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s}) \sum_{a\in A} \hat{\pi}(a|\bar{s}) \log(\pi_{t+1}(a|\bar{s}) Z_t(\bar{s})/\pi_t(a|\bar{s})) \\
&= \frac{1}{\zeta_t} \sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s}) (D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s})) - D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_{t+1}(\cdot|\bar{s})) + \log Z_t(\bar{s})) \\
&\overset{\text{(i)}}{\leq} \frac{1}{\zeta_t} \sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s}) (D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s})) - D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_{t+1}(\cdot|\bar{s}))) \\
&+ \frac{2\zeta_t}{(1-\gamma)^5} (R_{\max} + \zeta_t N \lambda_{\max} C_{\max})^2,
\end{aligned}
\tag{24}
$$

where (i) follows (23) by substituting ρ into $d_\rho^{\hat{\pi}}$. Now, let us consider the second case where constraints are violated. In this case, the policy is updated as follows:

$$
\theta_{t+1} = \theta_t + \frac{\zeta_t}{1-\gamma} (\zeta_t \nu_t A_R^t - \sum_i \lambda_{t,i} A_{i,g_i}^t). \tag{25}
$$

As (21) derived in the first case,

$$
\begin{aligned}
&\zeta_t\nu_t(J_R(\theta_{t+1}) - J_R(\theta_t)) + \sum_i \lambda_{t,i}(J_{C_i}(\theta_t) - J_{C_i}(\theta_{t+1})) \\
&= \frac{1}{1-\gamma}\sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_a \pi_{t+1}(a|\bar{s})(\zeta_t\nu_t A_R^t(\bar{s},a) - \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)) \\
&= \frac{1}{1-\gamma}\sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_a (\pi_{t+1}(a|\bar{s}) - \pi_t(a|\bar{s}))(\zeta_t\nu_t A_R^t(\bar{s},a) - \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)) \\
&\leq \frac{2}{1-\gamma}\sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \max_a \left|\zeta_t\nu_t A_R^t(\bar{s},a) - \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)\right| D_{\mathrm{TV}}(\pi_{t+1}(\cdot|\bar{s}), \pi_t(\cdot|\bar{s})) \\
&\leq \frac{2||\theta_{t+1} - \theta_t||_\infty}{1-\gamma}\sum_{\bar{s}} d_\rho^{\pi_{t+1}}(\bar{s}) \max_a \left|\zeta_t\nu_t A_R^t(\bar{s},a) - \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)\right| \\
&\leq \frac{2\zeta_t}{(1-\gamma)^2}\left|\left|\zeta_t\nu_t A_R^t - \sum_i \lambda_{t,i} A_{i,g_i}^t\right|\right|_\infty^2 \leq 2\zeta_t(\zeta_t\lambda_{\max}R_{\max} + N\lambda_{\max}C_{\max})^2/(1-\gamma)^4.
\end{aligned}
\tag{26}
$$

Using (18) and (25),

$$
\begin{aligned}
&\zeta_t\nu_t(J_R(\theta_{t+1}) - J_R(\theta_t)) + \sum_i \lambda_{t,i}(J_{C_i}(\theta_t) - J_{C_i}(\theta_{t+1})) \\
&= \frac{1}{1-\gamma}\sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_{a\in A} \pi_{t+1}(a|\bar{s})(\zeta_t\nu_t A_R^t(\bar{s},a) - \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)) \\
&= \frac{1}{\zeta_t}\sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) \sum_{a\in A} \pi_{t+1}(a|\bar{s}) \log(\pi_{t+1}(a|\bar{s})Z_t(\bar{s})/\pi_t(a|\bar{s})) \\
&= \frac{1}{\zeta_t}\sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s})(D_{\mathrm{KL}}(\pi_{t+1}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s})) + \log Z_t(\bar{s})) \\
&\geq \frac{1}{\zeta_t}\sum_{\bar{s}\in\bar{S}} d_\rho^{\pi_{t+1}}(\bar{s}) \log Z_t(\bar{s}) \geq \frac{1-\gamma}{\zeta_t}\sum_{\bar{s}\in\bar{S}} \rho(\bar{s}) \log Z_t(\bar{s}).
\end{aligned}
\tag{27}
$$

By combining (26) and (27), the following is satisfied for any ρ:

$$
\frac{1-\gamma}{\zeta_t}\sum_{\bar{s}\in\bar{S}} \rho(\bar{s}) \log Z_t(\bar{s}) \leq \frac{2\zeta_t}{(1-\gamma)^4}(\zeta_t\lambda_{\max}R_{\max} + N\lambda_{\max}C_{\max})^2. \tag{28}
$$

As in (24), the following is satisfied for any policy $\hat{\pi}$:

$$
\begin{aligned}
&\zeta_t\nu_t(J_R(\hat{\pi}) - J_R(\pi_t)) + \sum_i \lambda_{t,i}(J_{C_i}(\pi_t) - J_{C_i}(\hat{\pi})) \\
&= \frac{1}{1-\gamma}\sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s}) \sum_{a\in A} \hat{\pi}(a|\bar{s})(\zeta_t\nu_t A_R^t(\bar{s},a) - \sum_i \lambda_{t,i} A_{i,g_i}^t(\bar{s},a)) \\
&= \frac{1}{\zeta_t}\sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s}) \sum_{a\in A} \hat{\pi}(a|\bar{s}) \log(\pi_{t+1}(a|\bar{s})Z_t(\bar{s})/\pi_t(a|\bar{s})) \\
&= \frac{1}{\zeta_t}\sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s})(D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s})) - D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_{t+1}(\cdot|\bar{s})) + \log Z_t(\bar{s})) \\
&\leq \frac{1}{\zeta_t}\sum_{\bar{s}\in\bar{S}} d_\rho^{\hat{\pi}}(\bar{s})(D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_t(\cdot|\bar{s})) - D_{\mathrm{KL}}(\hat{\pi}(\cdot|\bar{s})||\pi_{t+1}(\cdot|\bar{s}))) \\
&+ \frac{2\zeta_t}{(1-\gamma)^5}(\zeta_t\lambda_{\max}R_{\max} + N\lambda_{\max}C_{\max})^2.
\end{aligned}
\tag{29}
$$

Now, the main inequalities for both cases have been derived. Let us denote by $\mathcal{N}$ the set of time steps in which the constraints are not violated. This means that if $t \in \mathcal{N}$, $J_{C_i}(\pi_t) \leq d_i \ \forall i$. Then, using

(24) and (29), we have:

$$\begin{aligned}
&\sum_{t\in\mathcal{N}}\left(\zeta_t(J_R(\hat{\pi})-J_R(\pi_t))-\zeta_t^2\sum_i\lambda_{t,i}(J_{C_i}(\hat{\pi})-J_{C_i}(\pi_t))\right)\\
&\quad+\sum_{t\notin\mathcal{N}}\left(\zeta_t^2\nu_t(J_R(\hat{\pi})-J_R(\pi))-\zeta_t\sum_i\lambda_{t,i}(J_{C_i}(\hat{\pi})-J_{C_i}(\pi_t))\right)\\
&\leq\sum_{\bar{s}\in\bar{S}}d_\rho^{\hat{\pi}}(\bar{s})D_{\mathrm{KL}}(\pi_f(\cdot|\bar{s})||\pi_0(\cdot|\bar{s}))+\sum_{t\in\mathcal{N}}\frac{2\zeta_t^2}{(1-\gamma)^5}(R_{\max}+\zeta_t N\lambda_{\max}C_{\max})^2\\
&\quad+\sum_{t\notin\mathcal{N}}\frac{2\zeta_t^2}{(1-\gamma)^5}(\zeta_t\lambda_{\max}R_{\max}+N\lambda_{\max}C_{\max})^2.
\end{aligned}\tag{30}$$

If $t\notin\mathcal{N}$ and $\lambda_{t,i}>0$, $J_{C_i}(\pi_t)>d_i$ and $\sum_i\lambda_{t,i}=1$. Thus, $\sum_i\lambda_{t,i}(J_{C_i}(\pi_f)-J_{C_i}(\pi_t))\leq-\eta$. By substituting $\hat{\pi}$ into π_f in (30),

$$\begin{aligned}
&\sum_{t\in\mathcal{N}}\left(\zeta_t(J_R(\pi_f)-J_R(\pi_t))-\zeta_t^2\sum_i\lambda_{t,i}(J_{C_i}(\pi_f)-J_{C_i}(\pi_t))\right)\\
&\quad+\sum_{t\notin\mathcal{N}}(\zeta_t\eta+\zeta_t^2\nu_t(J_R(\pi_f)-J_R(\pi)))\\
&\leq\sum_{\bar{s}\in\bar{S}}d_\rho^{\pi_f}(\bar{s})D_{\mathrm{KL}}(\pi_f(\cdot|\bar{s})||\pi_0(\cdot|\bar{s}))\\
&\quad+\sum_t\zeta_t^2\underbrace{\frac{2\max(\zeta_0,1)^2}{(1-\gamma)^5}(\max(\lambda_{\max},1)R_{\max}+N\lambda_{\max}C_{\max})^2}_{=:K_1}.
\end{aligned}$$

For brevity, we denote $D_{\mathrm{KL}}(\pi_f||\pi_0)=\sum_{\bar{s}\in\bar{S}}d_\rho^{\pi_f}(\bar{s})D_{\mathrm{KL}}(\pi_f(\cdot|\bar{s})||\pi_0(\cdot|\bar{s}))$. Since $\sum_{t\notin\mathcal{N}}\zeta_t=\sum_t\zeta_t-\sum_{t\in\mathcal{N}}\zeta_t$, (30) can be modified as follows:

$$\begin{aligned}
&\sum_{t\in\mathcal{N}}\zeta_t(J_R(\pi_f)-J_R(\pi_t)-\eta)+\eta\sum_t\zeta_t\\
&\leq D_{\mathrm{KL}}(\pi_f||\pi_0)+K_1\sum_t\zeta_t^2+\sum_{t\in\mathcal{N}}\left(\zeta_t^2\sum_i\lambda_{t,i}(J_{C_i}(\pi_f)-J_{C_i}(\pi_t))\right)\\
&\quad-\sum_{t\notin\mathcal{N}}\left(\zeta_t^2\nu_t(J_R(\pi_f)-J_R(\pi))\right)\\
&\leq D_{\mathrm{KL}}(\pi_f||\pi_0)+\sum_t\zeta_t^2\underbrace{\left(K_1+2N\lambda_{\max}\frac{C_{\max}}{1-\gamma}+2\lambda_{\max}\frac{R_{\max}}{1-\gamma}\right)}_{=:K_2}.
\end{aligned}$$

It can be simplified as follows:

$$\sum_{t\in\mathcal{N}}\zeta_t(J_R(\pi_f)-J_R(\pi_t)-\eta)+\eta\sum_t\zeta_t\leq D_{\mathrm{KL}}(\pi_f||\pi_0)+K_2\sum_t\zeta_t^2.$$

If π_0 is set with positive values in all actions, $D_{\mathrm{KL}}(\hat{\pi}||\pi_0)$ is bounded for $\forall\hat{\pi}$. Additionally, due to the Robbins-Monro condition, RHS converges to some real number as T goes to infinity. Since $\eta\sum_t\zeta_t$ in LHS goes to infinity, $\sum_{t\in\mathcal{N}}\zeta_t(J_R(\pi_f)-J_R(\pi_t)-\eta)$ must go to minus infinity. It means that

$\sum_{t\in\mathcal{N}}\zeta_t = \infty$ since $|J_R(\pi)| \le R_{\max}/(1-\gamma)$. By substituting $\hat{\pi}$ into π^* in (30),

$$\begin{aligned}
&\sum_{t\in\mathcal{N}}\zeta_t(J_R(\pi^*) - J_R(\pi_t)) - \sum_{t\notin\mathcal{N}}\zeta_t\left(\sum_i \lambda_{t,i}(J_{C_i}(\pi^*) - J_{C_i}(\pi_t))\right)\\
&\le D_{\mathrm{KL}}(\pi^*||\pi_0) + \sum_{t\in\mathcal{N}}\frac{2\zeta_t^2}{(1-\gamma)^5}(R_{\max} + \zeta_t N\lambda_{\max}C_{\max})^2\\
&\quad + \sum_{t\notin\mathcal{N}}\frac{2\zeta_t^2}{(1-\gamma)^5}(\zeta_t\lambda_{\max}R_{\max} + N\lambda_{\max}C_{\max})^2\\
&\quad + \sum_{t\in\mathcal{N}}\zeta_t^2\sum_i\lambda_{t,i}(J_{C_i}(\pi^*) - J_{C_i}(\pi_t)) - \sum_{t\notin\mathcal{N}}\zeta_t^2\nu_t(J_R(\pi^*) - J_R(\pi))\\
&\le D_{\mathrm{KL}}(\pi^*||\pi_0) + K_2\sum_t\zeta_t^2.
\end{aligned}$$

Since $\sum_i \lambda_{t,i}(J_{C_i}(\pi^*) - J_{C_i}(\pi_t)) \le 0$ for $t\notin\mathcal{N}$, the above equation is rewritten as follows:

$$\sum_{t\in\mathcal{N}}\zeta_t(J_R(\pi^*) - J_R(\pi_t)) \le D_{\mathrm{KL}}(\pi^*||\pi_0) + K_2\sum_t\zeta_t^2.$$

Since $\sum_{t\in\mathcal{N}}\zeta_t = \infty$ and the Robbins-Monro condition is satisfied, the following must be held:

$$J_R(\pi^*) - \lim_{t\to\infty} J_R(\pi_{\mathcal{N}[t]}) = 0,$$

where $\mathcal{N}[t]$ is the tth element of $\mathcal{N}$. Consequently, the policy converges to an optimal policy. □

A.2 Convergence Rate

In this section, we analyze the convergence rate of the proposed method. We adopt the finite-time analysis done by Xu et al. [25], and to do this, we make a few minor modifications to the policy update rule described in Section 5.2 as follows:

$$\theta_{t+1} = \begin{cases} \theta_t + \zeta F^\dagger(\theta_t)\nabla_\theta\left(J_R(\theta) - \zeta\sum_{i=1}^N\lambda_{t,i}J_{C_i}(\theta)\right)\big|_{\theta=\theta_t} & \text{if } J_{C_i}(\pi_t)\le d_i+\eta\ \forall i,\\ \theta_t + \zeta F^\dagger(\theta_t)\nabla_\theta\left(\zeta\nu_t J_R(\theta) - \sum_{i=1}^N\lambda_{t,i}J_{C_i}(\theta)\right)\big|_{\theta=\theta_t} & \text{otherwise,}\end{cases}$$

where η is a positive number. Then, we can show the convergence rate of the proposed method.

Theorem A.3. *Let us set* $\zeta = (1-\gamma)^{2.5}/\sqrt{T}$ *and* $\eta = (2D+20K^2)/((1-\gamma)^{2.5}\sqrt{T})$*, where* $D := \sum_s d_\rho^{\pi^*}(s)D_{\mathrm{KL}}(\pi^*(\cdot|s)||\pi_0(\cdot|s))$ *and* $K := \max(N\lambda_{\max}C_{\max}, \lambda_{\max}R_{\max}, R_{\max}, 1)$*. Then, the followings are satisfied:*

$$\begin{aligned}
J_R(\pi^*) - \mathbb{E}_{t\sim\mathcal{N}}[J_R(\pi_t)] &\le (2D+20K^2)/((1-\gamma)^{2.5}\sqrt{T}),\\
\mathbb{E}_{t\sim\mathcal{N}}[J_{C_i}(\pi_t)] - d_i &\le (2D+20K^2)/((1-\gamma)^{2.5}\sqrt{T})\ \textit{for}\ \forall i.
\end{aligned}$$

Proof. By subtituting $|J_R(\pi^*) - J_R(\pi_t)| \le 2R_{\max}/(1-\gamma)$ and $|J_{C_i}(\pi^*) - J_{C_i}(\pi_t)| \le 2C_{\max}/(1-\gamma)$ into (30),

$$\begin{aligned}
&\sum_{t\in\mathcal{N}}(\zeta(J_R(\pi^*) - J_R(\pi_t)) - 2\zeta^2 N\lambda_{\max}C_{\max}/(1-\gamma))\\
&\qquad + \sum_{t\notin\mathcal{N}}(\zeta\eta - 2\zeta^2\lambda_{\max}R_{\max}/(1-\gamma))\\
&\le D + \sum_{t\in\mathcal{N}}2\zeta^2(R_{\max} + \zeta N\lambda_{\max}C_{\max})^2/(1-\gamma)^5\\
&\qquad + \sum_{t\notin\mathcal{N}}2\zeta^2(\zeta\lambda_{\max}R_{\max} + N\lambda_{\max}C_{\max})^2/(1-\gamma)^5.
\end{aligned}$$

Using the definition of K,

$$\sum_{t\in\mathcal{N}}\zeta(J_R(\pi^*) - J_R(\pi_t)) + \sum_{t\notin\mathcal{N}}\zeta\eta \le D + 2T\zeta^2K^2(\zeta+1)^2/(1-\gamma)^5 + 2T\zeta^2K/(1-\gamma).$$

Given that $D + 2T\zeta^2K^2(\zeta+1)^2/(1-\gamma)^5 + 2T\zeta^2K/(1-\gamma) \leq \zeta\eta T/2$, one of the following must hold: **1)** $|\mathcal{N}| \geq T/2$ or **2)** $\sum_{t\in\mathcal{N}}(J_R(\pi^*) - J_R(\pi_t)) \leq 0$. Let us assume that neither condition is satisfied. Then,

$$\begin{aligned}\frac{\zeta\eta T}{2} < (T - |\mathcal{N}|)\zeta\eta = \sum_{t\notin\mathcal{N}} \zeta\eta < \sum_{t\in\mathcal{N}} \zeta(J_R(\pi^*) - J_R(\pi_t)) + \sum_{t\notin\mathcal{N}} \zeta\eta \\ \leq D + \frac{2T\zeta^2K^2(\zeta+1)^2}{(1-\gamma)^5} + \frac{2T\zeta^2K}{(1-\gamma)} \leq \frac{\zeta\eta T}{2},\end{aligned}$$

which leads to a contradiction. Therefore, one of the two conditions must hold. Furthermore, if we assume $|\mathcal{N}| = 0$, it follows that $\zeta\eta T \leq \zeta\eta T/2$, which is also a contradiction. Hence, $|\mathcal{N}| \neq 0$. If $|\mathcal{N}| \geq T/2$,

$$\begin{aligned}J_R(\pi^*) - \mathbb{E}_{t\sim\mathcal{N}}[J_R(\pi_t)] &= \frac{\sum_{t\in\mathcal{N}}(J_R(\pi^*) - J_R(\pi_t))}{|\mathcal{N}|} \\ &\leq \frac{2(D + \frac{2T\zeta^2K^2(\zeta+1)^2}{(1-\gamma)^5} + \frac{2T\zeta^2K}{(1-\gamma)})}{\zeta T} \\ &\leq \frac{2D + 20K^2}{(1-\gamma)^{2.5}\sqrt{T}}.\end{aligned}$$

The above inequality also holds for $\sum_{t\in\mathcal{N}}(J_R(\pi^*) - J_R(\pi_t)) \leq 0$. Consequently,

$$J_R(\pi^*) - \mathbb{E}_{t\sim\mathcal{N}}[J_R(\pi_t)] \leq (2D + 20K^2)/((1-\gamma)^{2.5}\sqrt{T}).$$

Also, $\mathbb{E}_{t\sim\mathcal{N}}[J_{C_i}(\pi_t)] - d_i \leq \eta = (2D + 20K^2)/((1-\gamma)^{2.5}\sqrt{T})$. □

B Weight Selection Strategy

In this section, we first identify ν_t and $\lambda_{t,i}$ for existing primal approach-based safe RL algorithms and present our strategy. For the existing methods, we analyze several safe RL algorithms including constrained policy optimization (CPO, [2]), projection-based constrained policy optimization (PCPO, [27]), constraint-rectified policy optimization (CRPO, [25]), penalized proximal policy optimization (P3O, [31]), interior-point policy optimization (IPO, [18]), and safe distributional actor-critic (SDAC, [15]). Since the policy gradient is obtained differently depending on whether the constraints are satisfied (CS) or violated (CV), we analyze the existing methods for each of the two cases.

Case 1: Constraints are satisfied (CS). First, CPO [2] and SDAC [15] find a direction of the policy gradient by solving the following linear programming with linear and quadratic constraints (LQCLP):

$$\max_g \nabla J_R(\theta_t)^T g \ \mathbf{s.t.} \nabla J_{C_i}(\theta_t)^T g + J_{C_i}(\theta_t) \leq d_i \ \forall i, \ \frac{1}{2} g^T F(\theta_t) g \leq \epsilon,$$

where ϵ is a trust region size. The LQCLP can be addressed by solving the following dual problem:

$$\begin{aligned}\lambda^*, \nu^* = \arg\max_{\lambda\geq 0, \nu\geq 0} -\nabla J_R(\theta_t)^T g^*(\lambda,\nu) + \sum_{i=1}^{N} \lambda_i(\nabla J_{C_i}(\theta_t)^T g^*(\lambda,\nu) + J_{C_i}(\theta_t) - d_i) \\ + \nu(\frac{1}{2} g^*(\lambda,\nu)^T F(\theta_t) g^*(\lambda,\nu) - \epsilon),\end{aligned}$$

where $g^*(\lambda,\nu) = \frac{1}{\nu} F^\dagger(\theta_t)\left(\nabla J_R(\theta_t) - \sum_i \lambda_i \nabla J_{C_i}(\theta_t)\right)$. Then, the policy is updated in the following direction:

$$g_t = F^\dagger(\theta_t)(\nabla J_R(\theta_t) - \zeta_t \sum_i \min(\lambda_i^*/\zeta_t, \lambda_{\max}) \nabla J_{C_i}(\theta_t)), \tag{31}$$

which results in $\lambda_{t,i} = \min(\lambda_i^*/\zeta_t, \lambda_{\max})$. P3O [31], PCPO [27], and CRPO [25] update the policy only using the objective function in this case as follows:

$$g_t = F^\dagger(\theta_t)\nabla J_R(\theta_t) \Rightarrow \lambda_{t,i} = 0.$$

IPO [18] update a policy in the following direction:

$$g_t = F^\dagger(\theta_t)\left(\nabla J_R(\theta_t) - \kappa \sum_i \nabla J_{C_i}(\theta_t)/(d_i - J_{C_i}(\theta_t))\right),$$

where κ is a penalty coefficient. As a result, $\lambda_{t,i}$ is computed as $\min(\kappa/((d_i - J_{C_i}(\theta_t))\zeta_t), \lambda_{\max})$.

Case 2: Constraints are violated (CV). CPO, PCPO and IPO did not handle the constraint violation in multiple constraint settings, so we exclude them in this case. First, SDAC finds a recovery direction by solving the following quadratic programming (QP):

$$\min_g \frac{1}{2} g^T F(\theta_t) g \text{ s.t.} \nabla J_{C_i}(\theta_t)^T g + J_{C_i}(\theta_t) \leq d_i \text{ for } i \in \{i | J_{C_i}(\theta_t) > d_i\},$$

where we will denote $\{i | J_{C_i}(\theta_t) > d_i\}$ by I_{CV}. The QP can be addressed by solving the following dual problem:

$$\lambda^* = \arg\max_{\lambda \geq 0} \frac{1}{2} g^*(\lambda)^T F(\theta_t) g^*(\lambda) + \sum_{i \in I_{\text{CV}}} \lambda_i (\nabla J_{C_i}(\theta_t)^T g^*(\lambda) + J_{C_i}(\theta_t) - d_i),$$

where $g^*(\lambda) = -F^\dagger(\theta_t) \sum_{i \in I_{\text{CV}}} \lambda_i \nabla J_{C_i}(\theta_t)$. Then, the policy is updated in the following direction:

$$g_t = F^\dagger(\theta_t) \left(\sum_{i \in I_{\text{CV}}} -\frac{\lambda_i^*}{\sum_{j \in I_{\text{CV}}} \lambda_j^*} \nabla J_{C_i}(\theta_t) \right) \Rightarrow \nu_t = 0,\ \lambda_{t,i} = \frac{\lambda_i^*}{\sum_{j \in I_{\text{CV}}} \lambda_j^*} \text{ if } i \in I_{\text{CV}} \text{ else } 0.$$

If there is only a single constraint, PCPO is compatible with SDAC. Next, CRPO first selects a violated constraint, whose index denoted as k, and calculates the policy gradient to minimize the selected constraint as follows:

$$g_t = F^\dagger(\theta_t) \left(-\nabla J_{C_k}(\theta_t) \right) \Rightarrow \nu_t = 0,\ \lambda_{t,i} = 1 \text{ if } i = k \text{ else } 0.$$

If there is only a single constraint, CPO is compatible with CRPO. Finally, P3O update the policy in the following direction:

$$g_t = F^\dagger(\theta_t) \left(\nabla J_R(\theta_t) - \kappa \sum_{i \in I_{\text{CV}}} \nabla J_{C_i}(\theta_t) \right),$$

which results in $\nu_t = \min(1/(\zeta_t \kappa |I_{\text{CV}}|), \lambda_{\max})$, and $\lambda_{t,i} = 1/|I_{\text{CV}}|$ if $i \in I_{\text{CV}}$ else 0.

Implementation of trust region method. To this point, we have obtained the direction of the policy gradient g_t, allowing the policy to be updated to $\theta_t + \zeta_t g_t$. If we want to update the policy using a trust region method, as introduced in TRPO [21], we can adjust the learning rate ζ_t as follows:

$$\zeta_t = \frac{\epsilon_t}{\text{Clip}(\sqrt{g_t^T F(\theta_t) g_t}, g_{\min}, g_{\max})}, \tag{32}$$

where ϵ_t is a trust region size following the Robbins-Monro condition, $g_{\min}$ and $g_{\max}$ are hyper-paramters, and $\text{Clip}(x, a, b) = \min(\max(x, a), b)$. By adjusting the learning rate, the policy can be updated mostly within the trust region, defined as $\{\theta_t + g | g^T F(\theta_t) g \leq \epsilon_t^2\}$, while the learning rate still satisfies the Robbins-Monro condition. Accordingly, we also need to modify $\lambda_{t,i}$ for the CS case and ν_t for the CV case, as they are expressed with the learning rate ζ_t. We explain this modification process by using SDAC as an example. The direction of the policy gradient of SDAC for the CS case expressed as follows:

$$g_t = F^\dagger(\theta_t)(\nabla J_R(\theta_t) - \epsilon_t \sum_i \min(\lambda_i^*/\epsilon_t, \lambda_{\max}) \nabla J_{C_i}(\theta_t)),$$

where ζ_t in (31) is replaced with ϵ_t. The learning rate ζ_t is then calculated using g_t and (32). Now, we need to express g_t as $F^\dagger(\theta_t)(\nabla J_R(\theta_t) - \zeta_t \sum_i \lambda_{t,i} \nabla J_{C_i}(\theta_t))$. To this end, we can use $\zeta_t = \epsilon_t / C$, where $C = \text{Clip}(\sqrt{g_t^T F(\theta_t) g_t}, g_{\min}, g_{\max})$. Consequently, $\lambda_{t,i}$ is expressed as follows:

$$\lambda_{t,i} = C \min(\lambda_i^*/\epsilon_t, \lambda_{\max}),$$

and the maximum of $\lambda_{t,i}$ and ν_t is adjusted to $g_{\max} \lambda_{\max}$.

Proposed strategy. Now, we introduce the proposed strategy for determining $\lambda_{t,i}$ and ν_t. We first consider the CV case. Kim et al. [15] empirically showed that processing multiple constraints simultaneously at each update step effectively projects the current policy onto a feasible set of policies.

Therefore, we decide to use the same strategy as SDAC, which deals with constraints simultaneously, and it is written as follows:

$$\nu_t = 0,\ \lambda_{t,i} = \begin{cases} \lambda_i^* / \left(\sum_{j \in I_{\mathrm{CV}}} \lambda_j^*\right) & \text{if } i \in I_{\mathrm{CV}}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda^* = \arg\max_{\lambda \geq 0} \frac{1}{2} g^*(\lambda)^T F(\theta_t) g^*(\lambda) + \sum_{i \in I_{\mathrm{CV}}} \lambda_i \left(\nabla J_{C_i}(\theta_t)^T g^*(\lambda) + J_{C_i}(\theta_t) - d_i\right)$ and $g^*(\lambda) = -F^\dagger(\theta_t) \sum_{i \in I_{\mathrm{CV}}} \lambda_i \nabla J_{C_i}(\theta_t)$. Next, let us consider the CS case. We hypothesize that concurrently addressing both constraints and objectives when calculating the policy gradient can lead to more stable constraint satisfaction. Therefore, CPO, SDAC, and IPO can be suitable candidates. However, CPO and SDAC methods require solving the LQCLP whose dual problem is nonlinear, and IPO can introduce computational errors when $J_{C_i}(\theta_t)$ is close to d_i. To address these challenges, we propose to calculate the policy gradient direction by solving a QP instead of an LQCLP. By applying a concept of transforming objectives into constraints [16] to the safe RL problem, we can formulate the following QP:

$$\min_g \frac{1}{2} g^T F(\theta_t) g \ \mathbf{s.t.}\ \nabla J_R(\theta_t)^T g \geq e,\ \nabla J_{C_i}(\theta_t)^T g + J_{C_i}(\theta_t) \leq d_i\ \forall i,$$

where $e := \epsilon_t \sqrt{\nabla J_R(\theta_t)^T F^\dagger(\theta_t) \nabla J_R(\theta_t)}$. According to [16], e represents the maximum value of $\nabla J_R(\theta_t)^T g$ when g is within the trust region, $\{g | g^T F(\theta_t) g \leq \epsilon_t^2\}$. Therefore, the transformed constraint, $\nabla J_R(\theta_t)^T g \geq e$, has the same effect of improving the objective function $J_R(\theta)$ within the trust region. The QP is then addressed by solving the following dual problem:

$$\begin{aligned} \lambda^*, \nu^* = \arg\max_{\lambda \geq 0, \nu \geq 0} & \frac{1}{2} g^*(\lambda, \nu)^T F(\theta_t) g^*(\lambda, \nu) + \nu(e - \nabla J_R(\theta_t)^T g^*(\lambda, \nu)) \\ & + \sum_{i=1}^{N} \lambda_i (\nabla J_{C_i}(\theta_t)^T g^*(\lambda, \nu) + J_{C_i}(\theta_t) - d_i), \end{aligned}$$

where $g^*(\lambda, \nu) = F^\dagger(\theta_t)(\nu \nabla J_R(\theta_t) - \sum_{i \in I_{\mathrm{CV}}} \lambda_i \nabla J_{C_i}(\theta_t))$. Then, the policy is updated in the following direction:

$$g_t = F^\dagger(\theta_t) \left(\nabla J_R(\theta_t) - \epsilon_t \sum_i \min\left(\frac{\lambda_i^*}{\nu^* \epsilon_t}, \lambda_{\max} \right) \nabla J_{C_i}(\theta_t) \right),$$

and the learning rate is calculated using (32) to ensure that the policy is updated within the trust region. Finally, $\lambda_{t,i}$ is expressed as $C \min(\lambda_i^*/(\nu^* \epsilon_t), \lambda_{\max})$, where $C = \mathrm{Clip}(\sqrt{g_t^T F(\theta_t) g_t}, g_{\min}, g_{\max})$.

C Implementation Details

In this section, we present a practical algorithm that utilizes function approximators, such as neural networks, to tackle more complex tasks. The comprehensive procedure is presented in Algorithm 2.

First, it is necessary to train individual policies for various $\boldsymbol{\beta}$, so we model a policy network to be conditioned on $\boldsymbol{\beta}$, expressed as $\pi_\theta(a|\bar{s}; \boldsymbol{\beta})$. This network structure enables the implementation of policies with various behaviors by conditioning the policy on different $\boldsymbol{\beta}$ values. In order to train this $\boldsymbol{\beta}$-conditioned policy across a broad range of the $\boldsymbol{\beta}$ space, a ϵ-greedy strategy can be used. At the beginning of each episode, $\boldsymbol{\beta}$ is sampled uniformly with a probability of ϵ; otherwise, it is sampled using the sampler ξ_ϕ. In our implementation, we did not use the uniform sampling process by setting $\epsilon = 0$ to reduce fine-tuning efforts, but it can still be used to enhance practical performance.

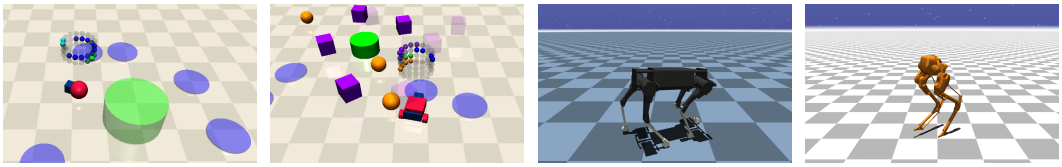
After collecting rollouts with the sampled $\boldsymbol{\beta}$, the critic, policy, and sampler need to be updated. For the critic, a distributional target is calculated using the TD(λ) method [15], and we update the critic networks to minimize the quantile regression loss [11], defined as follows:

$$\mathcal{L}(\psi) := \sum_{l=1}^{L} \mathop{\mathbb{E}}_{(\bar{s}, a, \boldsymbol{\beta}) \sim D} \left[\mathop{\mathbb{E}}_{z \sim \hat{Z}_{C_i}(\bar{s}, a; \boldsymbol{\beta})} \left[\rho_{\frac{2l-1}{2L}} \left(z - Z_{C_i, \psi}(\bar{s}, a; \boldsymbol{\beta})[l] \right) \right] \right],$$

where $\rho_l(u) := u \cdot (l - \mathbf{1}_{u<0})$, and $\hat{Z}_{C_i}(\bar{s}, a; \boldsymbol{\beta})$ is the target distribution of $Z_{C_i, \psi}(\bar{s}, a; \boldsymbol{\beta})$. The policy gradient for a specific $\boldsymbol{\beta}$ can be calculated according to the update rule described in Section

Algorithm 2 Spectral-Risk-Constrained Policy Optimization (SRCPO)

Input: Policy parameter θ, sampler parameter ϕ, critic parameter ψ, and replay buffer D.
for epochs $= 1$ **to** P **do**
  for episodes $= 1$ **to** E **do**
    if $u \leq \epsilon$, where $u \sim U[0, 1]$, **then**
      Sample $\boldsymbol{\beta}$ from a uniform distribution: $\beta_i[j] \sim U(\beta_i[j-1], C_{\max}/(1-\gamma))\ \forall i$ and $\forall j$, where $\beta_i[0] = 0$.
    else
      Sample $\boldsymbol{\beta} \sim \xi_\phi$.
    end if
    Collect a trajectory $\tau = \{\bar{s}_t, a_t, r_t, \{c_{i,t}\}_{i=1}^N\}_{t=0}^T$ using $\pi_\theta(a|\bar{s}; \boldsymbol{\beta})$, and store $(\boldsymbol{\beta}, \tau)$ in D.
  end for
  Set $G^\theta = \{\}$ and $G^\phi = \{\}$.
  for updates $= 1$ **to** U **do**
    Sample $(\boldsymbol{\beta}, \tau) \sim D$.
    Update the critics, $Z_{R,\psi}(\bar{s}, a; \boldsymbol{\beta})$ and $Z_{C_i,\psi}(\bar{s}, a; \boldsymbol{\beta})$, using the quantile regression loss.
    Calculate the policy gradient of $\pi_\theta(a|\bar{s}; \boldsymbol{\beta})$ using the update rule, described in Section 5.2, and store it to G^θ.
    Calculate the gradient of the sampler ξ_ϕ using the update rule, described in (14), and store it to G^ϕ.
  end for
  Update the policy parameter by the average of G^θ.
  Update the sampler parameter by the average of G^ϕ.
end for
Output: $\pi_\theta(\cdot|\cdot; \boldsymbol{\beta})$, where $\boldsymbol{\beta} \sim \xi_\phi$.



(a) Point goal task (b) Car button task (c) Quadrupedal robot (d) Bipedal robot

Figure 5: Rendered images of the Safety Gymnasium and the legged robot locomotion tasks.

5.2. Consequently, the policy is updated using the policy gradient calculated with $\boldsymbol{\beta}$ and rollouts, which are extracted from the replay buffer. Finally, the gradient of the sampler for a given $\boldsymbol{\beta}$ can be calculated as in (14), so the sampler is updated with the average of gradients, which are calculated across multiple $\boldsymbol{\beta}$.

D Experimental Details

In this section, we provide details on tasks, network structure, and hyperparameter settings.

D.1 Task Description

There are two environments, legged robot locomotion [15] and Safety Gymnasium [14], and whose rendered images are presented in Figure 5.

Legged Robot Locomotion. The legged robot locomotion tasks aim to control bipedal and quadrupedal robots to maintain a specified target velocity, which consists of linear velocity in the X and Y directions and rotational velocity in the Z direction, without falling. Each robot takes actions as PD target joint positions at a frequency of 50 Hz, and a PD controller in each joint operates at 500 Hz. The dimension of the action space corresponds to the number of motors: ten for the bipedal robot and 12 for the quadrupedal robot. The state space includes the command, gravity vector, current joint positions and velocities, foot contact phase, and a history of the joint positions and

velocities. The reward function is defined as the negative of the squared Euclidean distance between the current velocity and the target velocity. There are three cost functions to prevent robots from falling. The first cost is to balance the robot, which is formulated as $\mathbf{1}_{g[2]/||g||_2 \geq \cos(\pi/12)}$, where g is the gravity vector. This function is activated when the base frame of the robot is tilted more than 15 degrees. The second cost is to maintain the height of the robot above a predefined threshold, which is expressed as $\mathbf{1}_{h<b}$, where h is the height of the robot, and b is set as 0.7 in the bipedal robot and 0.35 in the quadrupedal robot. The third cost is to match the foot contact state with the desired timing. It is defined as the average number of feet that do not touch the floor at the desired time. We set the thresholds for these cost functions as $0.025/(1-\gamma), 0.025/(1-\gamma)$, and $0.4/(1-\gamma)$, respectively, where $(1-\gamma)$ converts the threshold value from an average time horizon scale to an infinite time horizon scale.

Safety Gymnasium. We use the goal and button tasks in the Safety Gymnasium [14], and both tasks employ point and car robots. The goal task is to control the robot to navigate to a target goal position without passing hazard area. The button task is to control the robot to push a designated button among several buttons without passing hazard area or hitting undesignated buttons. The dimension of the action space is two for all tasks. The state space dimensions are 60, 72, 76, and 88 for the point goal, car goal, point button, and car button tasks, respectively. For more details on the action and state, please refer to [14]. The reward function is defined to reduce the distance between the robot and either the goal or the designated button, with a bonus awarded at the moment of task completion. When the robot traverses a hazard area or collides with an obstacle, it incurs a cost of one, and the threshold for this cost is set as $0.025/(1-\gamma)$.

D.2 Network Structure

Across all algorithms, we use the same network architecture for the policy and critics. The policy network is structured to output the mean and standard deviation of a normal distribution, which is then squashed using the Tanh function as done in SAC [13]. The critic network is based on the quantile distributional critic [11]. Details of these structures are presented in Table 1. For the sampler ξ_ϕ, we define a trainable variable $\phi \in \mathbb{R}^{N\times(M-1)}$ and represent the mean of the truncated normal distribution as $\mu_{i,\phi}[j] = \exp(\phi[i,j])$. To reduce fine-tuning efforts, we fix the standard deviation of the truncated normal distribution to 0.05.

Table 1: Details of network structures.

	Parameter	Value
Policy network	Hidden layer	(512, 512)
	Activation	LeakyReLU
	Last activation	Linear
Reward critic network	Hidden layer	(512, 512)
	Activation	LeakyReLU
	Last activation	Linear
	# of quantiles	25
	# of ensembles	2
Cost critic network	Hidden layer	(512, 512)
	Activation	LeakyReLU
	Last activation	SoftPlus
	# of quantiles	25
	# of ensembles	2

D.3 Hyperparameter Settings

The hyperparameter settings for all algorithms applied to all tasks are summarized in Table 2.

Table 2: Description on hyperparameter settings.

	SRCPO (Ours)	CPPO [29]	CVaR-CPO [32]	SDAC [15]	SDPO [30]	WCSAC-D [26]
Discount factor	0.99	0.99	0.99	0.99	0.99	0.99
Policy learning rate	-	$3 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	-	$3 \cdot 10^{-5}$	$3 \cdot 10^{-4}$
Critic learning rate	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$
# of target quantiles	50	50	50	50	50	-
Coefficient of TD(λ)	0.97	0.97	0.97	0.97	0.97	-
Soft update ratio	-	-	-	-	-	0.995
Batch size	10000	5000	5000	10000	5000	256
Steps per update	1000	5000	5000	1000	5000	100
# of policy update iterations	10	20	20	-	20	10
# of critic update iterations	40	40	40	40	40	10
Trust region size	0.001	0.001	0.001	0.001	0.001	-
Line search decay	-	-	-	0.8	-	-
PPO Clip ratio [22]	-	0.2	0.2	-	0.2	-
Length of replay buffer	100000	5000	5000	100000	5000	1000000
Learning rate for auxiliary variables	-	0.05	0.01	-	-	-
Learning rate for Lagrange multipliers	-	0.05	0.01	-	-	0.001
Log penalty coefficient	-	-	-	-	5.0	-
Entropy threshold	-	-	-	-	-	$-\|A\|$
Entropy learning rate	-	-	-	-	-	0.001
Entropy coefficient		0.001	-	-	0.001	-
Sampler learning rate	$1 \cdot 10^{-3}$	-	-	-	-	-
K for sampler target	10	-	-	-	-	-

E Additional Experimental Results

E.1 Computational Resources

All experiments were conducted on a PC whose CPU and GPU are an Intel Xeon E5-2680 and NVIDIA TITAN Xp, respectively. The training time for each algorithm on the point goal task is reported in Table 3.

Table 3: Training time for the point goal task averaged over five runs.

	SRCPO (Ours)	CPPO	CVaR-CPO	SDAC	WCSAC-D	SDPO
Averaged training time	9h 51m 54s	6h 17m 16s	4h 43m 38s	8h 46m 8s	14h 18m 20s	6h 50m 39s

E.2 Experiments on Various Risk Measures

We have conducted studies on various risk measures in the main text. In this section, we first describe the definition of the Wang risk measure [24], show the visualization of the discretization process, and finally present the training curves of the experiments conducted in Section 8.1.

The Wang risk measure is defined as follows [24]:

$$\text{Wang}_\alpha(X) := \int_0^1 F_X^{-1}(u) d\chi_\alpha(u), \quad \text{where } \chi_\alpha(u) := \Phi(\Phi^{-1}(u) - \alpha),$$

Φ is the cumulative distribution function (CDF) of the standard normal distribution, and χ is called a distortion function. It can be expressed in the form of the spectral risk measure as follows:

$$\text{Wang}_\alpha(X) = \int_0^1 F_X^{-1}(u)\sigma_\alpha(u) du, \quad \text{where } \sigma_\alpha(u) := \frac{\phi(\Phi^{-1}(u) - \alpha)}{\phi(\Phi^{-1}(u))},$$

and ϕ is the pdf of the standard normal distribution. Since $\sigma_\alpha(u)$ goes to infinity when u is approaching one, $\sigma_\alpha(u)$ at $u = 1$ is not defined. Thus, it is not a spectral risk measure. Nevertheless, our

parameterization method introduced in Section 6.1 can project Wang_α onto a discretized class of spectral risk measures.

The discretization process can be implemented by minimizing (11), and the results are shown in Figure 6. In this process, the number of discretizations M is set to five. Additionally, visualizations of other risk measures are provided in Figure 7, and the optimized values of the parameters for the discretized spectrum, which are introduced in (10), are reported in the Table 4. Note that CVaR is already an element of the discretized class of spectral risk measures, as observed in Figure 7. Therefore, CVaR can be precisely expressed using a parameter with $M = 1$.

Finally, we conducted experiments on the point goal task to check how the results are changed according to the risk level and the type of risk measures. The training curves for these experiments are shown in Figure 8.

Table 4: Discretization results.

		Wang_α			Pow_α
$\alpha = 0.5$	$\{\eta_i\}_{i=1}^{M}$ $\{\alpha_i\}_{i=1}^{M-1}$	$[0.515, 0.790, 1.091, 1.493, 2.191]$ $[0.263, 0.541, 0.770, 0.926]$	$\alpha = 0.5$	$\{\eta_i\}_{i=1}^{M}$ $\{\alpha_i\}_{i=1}^{M-1}$	$[0.200, 0.600, 1.000, 1.400, 1.800]$ $[0.200, 0.400, 0.600, 0.800]$
$\alpha = 1.0$	$\{\eta_i\}_{i=1}^{M}$ $\{\alpha_i\}_{i=1}^{M-1}$	$[0.294, 0.734, 1.417, 2.640, 5.517]$ $[0.409, 0.701, 0.878, 0.968]$	$\alpha = 0.75$	$\{\eta_i\}_{i=1}^{M}$ $\{\alpha_i\}_{i=1}^{M-1}$	$[0.046, 0.574, 1.347, 2.308, 3.424]$ $[0.417, 0.615, 0.765, 0.890]$
$\alpha = 0.5$	$\{\eta_i\}_{i=1}^{M}$ $\{\alpha_i\}_{i=1}^{M-1}$	$[0.180, 0.834, 2.253, 5.678, 16.419]$ $[0.579, 0.834, 0.945, 0.989]$	$\alpha = 0.9$	$\{\eta_i\}_{i=1}^{M}$ $\{\alpha_i\}_{i=1}^{M-1}$	$[0.003, 0.947, 2.705, 5.216, 8.383]$ $[0.701, 0.821, 0.898, 0.955]$

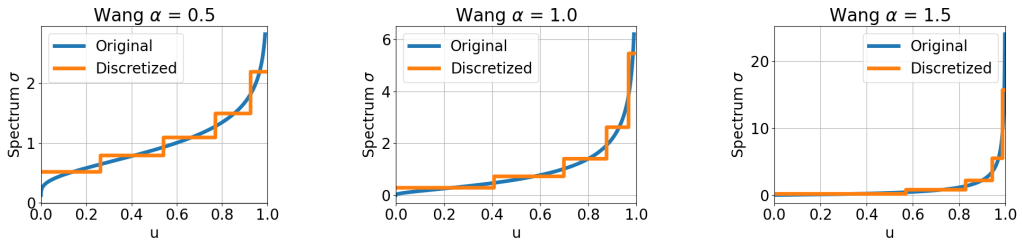


Figure 6: Visualization of the spectrum function of the Wang risk measure and the discretized results.

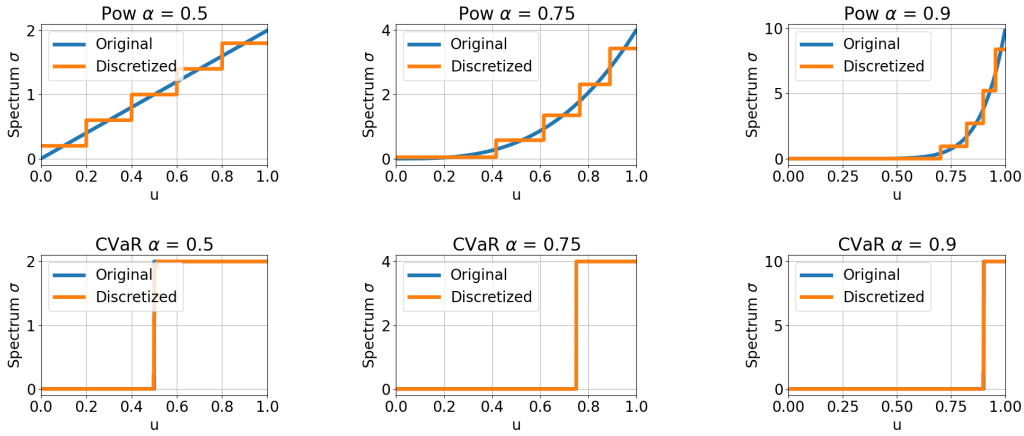


Figure 7: Visualization of the spectrum function of the Pow and CVaR risk measures and the discretized results.

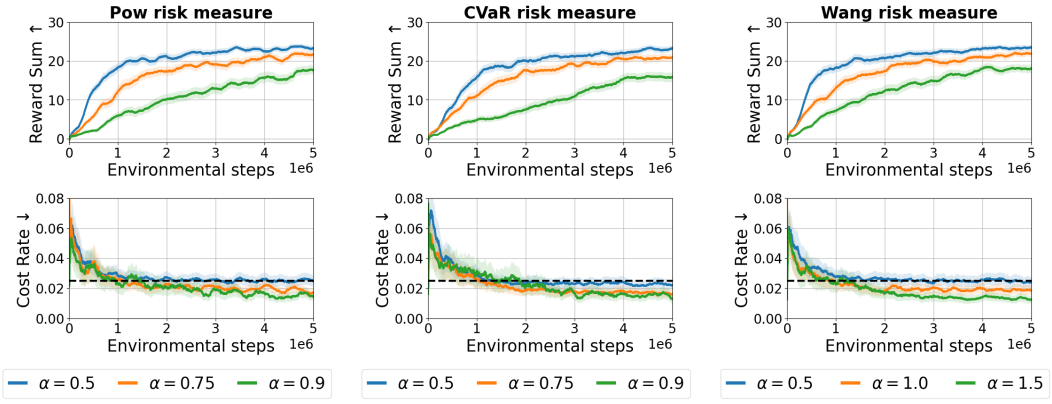


Figure 8: Training curves of the point goal task with different risk levels and risk measures. Each column shows the results on the risk measure whose name appears in the plot title. The solid line in each graph represents the average of each metric, and the shaded area indicates the standard deviation scaled by 0.2. The results are obtained by training algorithms with five random seeds.

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 main claims in the abstact and introduction reflect the contributions of 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 have discussed the limitations in the last section of the main text.

   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 have provided the full set of assumptions and a proof, which is presented in Section A.

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: We have enclosed the code and provided the pseudo-code for reproducibility in Appendix C.

   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 have enclosed the code.

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 setting and details are provided in Appendix D.

   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 have reported the standard deviation of experimental results.

   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: We have described the computational resources in Section E.

   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 paper conforms 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: [NA]

    Justification: This work has no societal impact.

    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: This work has no such risks.

    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 have cited the benchmark papers and included the licence for assets in the code.

    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: This work does 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: This works does not involve crowdsourcing nor research with human subjects.

    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: This work does not involve crowdsourcing nor research with human subjects.

    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.