ROSCOM: Robust Safe Reinforcement Learning on Stochastic Constraint Manifolds

Shangding Gu, Puze Liu, Alap Kshirsagar, Guang Chen, Jan Peters *Fellow, IEEE*, Alois Knoll *Fellow, IEEE*

***Abstract*—Reinforcement Learning (RL) has demonstrated remarkable success across various domains. Nonetheless, a significant challenge in RL is to ensure safety, particularly when deploying it in safety-critical applications such as robotics and autonomous driving. In this work, we develop a robust and safe RL methodology grounded in manifold space. Initially, we construct a constrained manifold space, taking safety constraints into consideration. We then propose a robust safe RL approach, supported by theoretical analysis, based on the value at risk and conditional value at risk, in order to enhance the robustness of safety. Our methodology is designed to ensure safety within stochastic constraint environments. Following the theoretical analysis, we develop a practical, safe algorithm to search for a robust safe policy on stochastic constraint manifolds (ROSCOM). We evaluate the effectiveness of our approach through circular motion and air-hockey tasks. Our experiments demonstrate that ROSCOM outperforms existing baselines in terms of both reward and safety.**

***Note to Practitioners*—Real-world applications often involve inherent uncertainties, noise, and high-dimensional spaces. This complexity accentuates the urgency and challenge of ensuring safety in robot learning, especially when implementing RL in practical environments. To address this critical issue, we build a stochastic constraint manifold to delineate the safety space, thus establishing a rigorous framework for robot learning at each iteration. Compared with state-of-the-art baselines, our method can provide remarkable performance regarding safety and reward performance. For example, in an air hockey robot learning task, our method has demonstrated a remarkable 50% enhancement in safety performance compared to the ATACOM framework [1], while concurrently exhibiting superior reward performance. Moreover, in contrast to traditional algorithms, including CPO [2], PCPO [3], our method has achieved a 99% improvement in safety performance, coupled with significantly superior reward performance. These empirical insights render our approach not only theoretically sound but also practically efficacious, indicating its potential as a useful tool in real robot learning and beyond.**

***Index Terms*—Safe Reinforcement Learning, Constrained Manifolds, Robust Reinforcement Learning, Robotics.**

Manuscript received October 26, 2023; revised February 14, 2024; accepted July 6, 2024. This work is supported by the National Natural Science Foundation of China (No. 62372329), in part by Shanghai Scientific Innovation Foundation (No.23DZ1203400), in part by Xiaomi Young Talents Program, and in part by "The Adaptive Mind", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art. (*Corresponding authors: Guang Chen.*)

Shangding Gu and Alois Knoll are with the Department of Informatics, Technical University of Munich, Munich 85748, Germany (e-mail: shangding.gu@tum.de; knoll@mytum.de).

Puze Liu, Alap Kshirsagar and Jan Peters are with the Department of Informatics, Technical University of Darmstadt, Darmstad 64289, Germany (e-mail: puze.liu@ias.tu-darmstadt.de; alap.kshirsagar@tu-darmstadt.de; jan.peters@tu-darmstadt.de).

Guang Chen is with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: guangchen@tongji.edu.cn).

I. Introduction

REINFORCEMENT Learning (RL) has garnered substantial attention in recent years due to its capability to address complex problems and exhibit impressive performance in various tasks [4], [5], such as AlphaGo [6], finance [7], multi-robot control [8], [9], and autonomous driving [10]. However, the deployment of RL in real-world applications, particularly in safety-critical systems, presents challenges as an RL agent may execute unsafe actions that could harm humans or damage the agent's environment [4]. Safe RL methods tackle this issue by incorporating safety requirements during the learning process and ensuring that the decisions made by the RL agent do not lead to hazardous outcomes.

Numerous state-of-the-art (SOTA) methods have been proposed to ensure RL safety, including CPO [2], PCPO [3], RCPO [11], MACPO [8], and MAPPO-Lagrangian [8]. However, these approaches are not applicable to situations where the robot constraints are stochastic. Such scenarios are frequently encountered in real-world applications with high-dimensional space, like robotics and autonomous driving. The robot's observations may exhibit uncertainty due to noisy sensors, while the robot's constraints may be stochastical owing to the presence of other agents, such as humans. Thus, to ensure the safety of a robot and its environment, it is essential to consider stochastic constraints during robot learning in high-dimensional space.

In this work, we aim to address the critical question: *How can we guarantee RL safety with stochastic constraints in high-dimensional space?* To ensure safety in stochastic environments with high-dimensional space, we need to calculate safety bounds and construct constraint manifolds that incorporate these bounds. Manifolds are highly effective for managing high-dimensional data, a fact that is well-supported by numerous studies [12]–[14]. With the constraint manifolds, we can search for a safe policy for robot learning, where the constraint manifold is well-suited for representing high-dimensional constraints. The contributions of this work are as follows:

- We introduce a problem formulation that considers stochastic constraints on manifolds. Specifically, the stochastic constraints are incorporated into the state space based on manifolds.
- We offer theoretical safety bounds on the stochastic constraint manifold by leveraging Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) methods [15], [16].

We propose an algorithm for searching a robust safe policy on a stochastic constraint manifold (ROSCOM) and demonstrate its effectiveness compared to several SOTA safe RL algorithms.

II. Related Work

Deploying RL in real-world applications necessitates the assurance of safety for both the agent and the environment, in addition to robustness in the face of uncertainties. Researchers have proposed a variety of safe and robust RL methods to apply RL in safety-critical environments, such as autonomous driving and robotics [4], [10], [17], [18]. Despite these efforts, there remains a need for further development of RL methods that can address safety and robustness challenges concurrently. This section provides a brief analysis of the current state of research in the domain of safe and robust RL.

A. Safe Reinforcement Learning.

In recent years, research on safe RL has gained significant attention due to its importance in addressing the safety concerns in various applications [4]. Safe RL can be formulated as a constrained optimization problem with methods falling into three main categories: constrained state space, constrained action space, and constrained cumulative cost.

The first category comprises methods that leverage constrained state space as a safe state space [19]–[22]. These methods often incorporate Gaussian Process models to estimate safe state space during exploration [19]–[22]. Another notable approach is ATACOM [1], which projects exploration states onto a constrained manifold. Our method also belongs to this category.

The second category includes methods that employ a constrained action space as a safe action space [23]–[28]. For instance, some methods use temporal logic verification to ensure the safety of actions during exploration [26], while others rely on Lyapunov functions to constrain the action space with energy functions [23]–[25].

The third category focuses on optimizing safe policies based on cumulative safety cost [2], [3], [8], [11]. Examples include CPO [2] for single-agent settings and MACPO [8] and MAPPO-Lagrangian [8] for multi-agent settings. These methods demonstrate the versatility of Safe RL approaches in addressing safety concerns across various domains and settings.

B. Robust Reinforcement Learning.

Existing robust RL methods tackle three types of uncertainties: reward uncertainty [29], [30], transition uncertainty [31]–[38], and observation uncertainty [39]–[43].

To address reward uncertainty, Liang *et al.* [29] proposed a human-preferences-based learning method, while Lee *et al.* [30] developed a meta-learning approach that learns uncertain rewards using normalized maximum likelihood.

For transition uncertainty, Chua *et al.* [31] introduced a probabilistic dynamics model to balance the trade-off between model-based and model-free RL methods. Zhang *et al.* [32] employed a natural player to disrupt the multi-agent system's transitions, similar to an adversarial agent, enabling the ego player to enhance its performance in adversarial environments.

Regarding observation uncertainty, Lutjens *et al.* [39] presented a certified adversarial robustness method for handling uncertainties during exploration. Zhang *et al.* [41] proposed a robust policy regularizer for uncertain observations in both discrete and continuous environments. These methods demonstrate the breadth of approaches in robust RL for addressing various uncertainties, thus facilitating the development of more reliable and resilient agents in diverse application scenarios.

The existing methods primarily focus on addressing either the safety or robustness aspect of RL, but not both simultaneously. The challenge of robust and safe RL lies in ensuring safety in uncertain environments, such as those with uncertain observations and constraints. The work of Liu *et al.* [18] is the closest to addressing this challenge, as they proposed a robust and safe RL method in an adversarial setting where two attackers disturb agent behaviors. However, the adversarial setting may limit the practicality of their method in real-world applications.

In contrast to Liu *et al.* [18], our approach ensures RL safety under stochastic constraints by employing a constraint manifold, which is capable of handling high-dimensional data in real-world environments. The stochastic constraints are generated by a stochastic function rather than an adversarial agent. Unlike Liu *et al.* [18], who employ neural networks to model the adversarial agent's behaviors. Lastly, our work provides theoretical safety bounds based on the constraint manifold, ensuring safety in stochastic constraints.

Our method focuses on the observation uncertainty, where the state space is disturbed by some noise, and we need to ensure safety via safety bounds. By addressing both safety and robustness in RL simultaneously, our method paves the way for more reliable and resilient agents that can operate effectively in safety-critical applications and stochastic constrained environments.

III. Problem Formulation

A robust safe RL problem [4] can be seen as a standard MDP with a constraint set under stochastic constraints, to illustrate this problem, we consider a tuple $(S, A, P, r, \rho_0, \gamma, c_i, \epsilon_i)$, where S is a state space, A is an action space, $P : S \times A \times S \to [0, 1]$ is a transition function, r denotes the reward, ρ_0 is the initial state distribution, γ is a discount factor, $c_i : S \to \mathbb{R}$ denotes a deterministic state-constraint function of constraint i, $\forall i = 1, \ldots, N$, $\epsilon_i : S \to \mathbb{R}$ denotes a stochastic state-constraint function of constraint i.

In robust safe RL, the goal of an optimal policy is to maximize the reward while ensuring safety by selecting an action a under stochastic constraints. τ is a trajectory, $\tau = (s_0, a_0, s_1, \ldots)$, which depends on π, $s_0 \sim \rho_0$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim P$, where the state $s_t \in S$ at time step t, the action $a_t \in A$ at time step t, $t \in \{0, 1, \ldots, T\}$, T is the episode length.

We formulate the robust safe RL problem as an MDP problem as shown in Equation (1), in which the stochastic

(a) (b)

Fig. 1: Schematic diagram of stochastic constraints and safety bound. (a). 3D stochastic constraints, Gaussian noise in the x, y, and z directions; (b). 3D safety bound, the areas of dark brown and red denote the new safe space, and the area of light brown denotes the uncertain space.

constraints are taken into account. Figure 1 shows an example of the stochastic constraints. In this Equation, we need to search for a policy to maximize the discounted cumulative reward value while ensuring learning safety. Speci cally, the safety constraints consist of deterministic constraints $c_i(s_t)$ and stochastic constraints $_i(s_t)$ at time step t, the stochastic constraints $_i(s_t)$ are from some uncertain noises due to imperfect observation. The aggregate of these constraints, $c_i(s_t)$ plus $_i(s_t)$, need to remain below a prede ned threshold with a probability of . Particularly, for all states s_t, if the constraints $c_i(s_t)$ and $_i(s_t)$ are equal to zero, then states s_t are the safe set (safe set indicates that we can ensure robot learning safety within these states), that is the safe set $= \{s_t \mid c_i(s_t) = 0;\ _i(s_t) = 0\}$. Conversely, for all states within the unsafe set, $c_i(s_t)$ and $_i(s_t)$ are positive real values, expressed as R^+. That is the unsafe set $= \{s_t \mid c_i(s_t) \in R^+;\ _i(s_t) \in R^+\}$. For each type of constraint, there corresponds a safety bound $_i$ and a safety probability $_i$. For simplicity, we denote these as and , respectively.

$$\max \ E\left[\sum_t {}^t r(s_t, a_t)\right];$$
$$\text{s.t. } Pr(|c_i(s_t) + {}_i(s_t)| \quad) > : \tag{1}$$

IV. METHOD

We rst present the de nitions of state constraint manifolds. Subsequently, we provide the theoretical safety bounds based on VaR and CVaR to search for safe policies. Sections IV-A is to build a stochastic constrained manifold, and Section IV-B aims to provide safety bounds on a stochastic constrained Manifold. Lastly, Section IV-C presents a practical algorithm for searching robust safe policies within a stochastic constraint manifold.

A. State Constraints on a Manifold

Similiar to ATACOM [1], in the high-dimensional state space S (the state variables $s \in S$) is consisted of two sets, a controllable set $q \in Q \subseteq R^Q$ and an uncontrollable set $x \in X$, $s = [q; x]^\top$. In this study, constraints are de ned on the controllable set, $Pr(|c_i(q) + {}_i(q)| <) > $, where $c_i = [f_i; g_i] \in R^{F+G}$ denotes constraints, equality constraints $f_i = 0$; $f_i \in R^F$ and inequality constraints $g_i < 0$; $g_i \in R^G$, $_i \in R^{F+G}$ denotes stochastic constraints, $\in R^{F+G}$ denotes a set of safety bounds, for simplicity, we use $\in R^{F+G}$ to denote a set of safety probability $^{;o} \in \{^{;1}, ^{;2}, \ldots, ^{;F+G}\}$, f_i to denotes a set of equality constraints $\{^{_1;o_1} \in \{^{1;1}, ^{1;2}, \ldots, ^{1;F}\}$, g_i to denotes a set of equality constraints $\{^{_2;o_2} \in \{^{2;1}, ^{2;2}, \ldots, ^{2;G}\}$. Note, when comparing values of a high-dimensional data set, we assess each individual element within the data set.

De nition 1. If the constraints are differentiable, we can have the following constraint manifolds: M_1 is the equality constraints manifold with stochastic constraints, M_2 is the inequality constraints manifold with stochastic constraints, as shown in Equation (2):

$$\begin{aligned} M_1 &= \{q \mid f_i(q) + {}_{f_i}(q) \quad = 0\} \\ M_2 &= \{q \mid g_i(q) + {}_{g_i}(q) \quad + \tfrac{1}{2} {}^2_s \geq 0\} \end{aligned} \Longrightarrow$$
$$\begin{aligned} M_{1V} &= \{f_{iV}(q) = 0\} \\ M_{2V} &= \{g_{iV}(q) + \tfrac{1}{2} {}^2 \geq 0\} \end{aligned} \tag{2}$$

In Equation (2), M_{1V} and M_{2V} represent the manifolds of equality and inequality constraints with VaR (Value at Risk) bounds, respectively. In the following section, we will describe how to compute these manifolds to ensure safety. $\in R^F$ denotes the stochastic constraints on the equality constraint manifold, $_{g_i} \in R^G$ denotes the stochastic constraints on the inequality constraint manifold, $f_{iV}(q)$ and $g_{iV}(q)$ denotes equality constraints and inequality constraints with VaR safety bounds considering stochastic constraints. To construct the safety manifold under inequality constraints, the slack variables $_s \in R^G$ and $\in R^G$ are leveraged, which are introduced into the constraints [1].

Therefore, we have the new constraint set $c(q; _s) \in R^{F+G}$, as shown in Equation (3). $J_{fiV}(q) \in R^{F \times Q}$ is the Jacobian matrice of $_{iV}(q) \in R^F$. We have two equality constraint sides when inequality constraints are converted to equality constraints with safety bounds. Thus, we consider two safety sides of the equality constraints $J^+_{fiV}(q) \in R^{F \times Q}$ and $J_{fiV}(q) \in R^{F \times Q}$. $J^+_{fiV}(q) \in R^{F \times Q}$ and $J_{fiV}(q) \in R^{F \times Q}$ are the Jacobian matrices of the safety bounds for the function $g_{iV}(q) \in R^G$.

$$c(q; \ ; q; \_) = \begin{bmatrix} J_{fiV}(q) + & 0 \\ J_{fiV}(q) & 0 \\ J_{giV}(q) & \mathrm{diag}(\ _s) \end{bmatrix} \begin{matrix} q \\ \_s \end{matrix} \tag{3}$$
$$= J_c(q; \ _s) \begin{matrix} q \\ \_s \end{matrix}; \tag{4}$$

where $J_c(q; \ _s) \in R^{(2F+G) \times (Q+G)}$ is the Jacobian Matrix, by leveraging SVD [44] and QR [45], we can have $J_c(q; \ _s) N_c(q; \ _s) = 0$, $N_c(q; \ _s) \in R^{(Q+G) \times (Q-2F)}$ is the null space, which can be seen as the tangent space bases of a constraint manifold.

Remark 1. Inspired by ATACOM [1], the nullspace of the Jacobian matrix $J_c(q; s) \in R^{(2F+G)\ (Q+G)}$ for the tangent-space bases of the constraint manifold is $N_c(q; s) \in R^{(Q+G)\ (Q\ 2F)}$, we can search a safe policy at each time step t on the constraint manifold $c(q; s)$.

In the subsequent section, we will delve into the methodologies and strategies to guarantee safety on a stochastic manifold. The complexities inherent to these manifolds necessitate rigorous approaches to maintain safety, and our discussion will shed light on these critical procedures.

B. Ensuring Safety with VaR and CVaR on a Stochastic Constraint Manifold

This section presents a theoretical analysis of the safety bounds based on stochastic constraints. First, we give the definition of the safety bounds and the risk measurements. Then, we derive safety bounds for stochastic constraints by leveraging VaR and CVaR [15], [16].

Definition 2. VaR and CVaR. For safety constraints $jc_i(q) + \ _i(q)j$, the value-at-risk (VaR) of $jc_i(q) + \ _i(q)j$ with confidence level $\in (0; 1)$ is defined as:

$$\text{VaR}\ (jc_i(q) + \ _i(q)j) = \min f\ \ j\ F(\ ) \qquad g;$$

where $F(\ ) = P(jc_i(q) + \ _i(q)j \qquad)$ is the cumulative distribution function (CDF), $_i(q)$ denotes the stochastic costs. The conditional value-at-risk (CVaR) of $jc_i(q) + \ _i(q)j$ with confidence level is defined as the expectation of the -tail distribution of $jc_i(q) + \ _i(q)j$ as

$$\text{CVaR}\ (\ _i(q)) = E f\ \ j \qquad \text{VaR}\ (jc_i(q) + \ _i(q)j)g:$$

Here, we assume stochastic constraints $_i(q)$ subject to a Gaussian distribution N, $Z = \ _i(q) \qquad N(\ ;\ ^2)$.

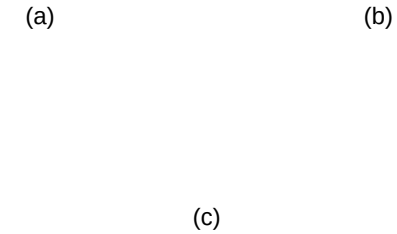
1) Safety Bound for Inequality Constraints with Stochastic Noises on a Manifold:

We rewrite VaR for inequality constraints as following equations, where $g_i(q)^a$ denotes the old deterministic inequality constraints, $g_i(q)$ denotes the new deterministic inequality constraints, $_g$ denotes the safety bound for inequality constraints, denotes the mean of stochastic constraints $_i(q)$, denotes the standard deviation of stochastic constraints $_i(q)$, is the safety probability for robot learning.

$$F(\ _g) = P(g_i(q)^a + \ _i(q) \qquad _g); \tag{5}$$

$$\begin{aligned}
&\text{VaR}\ (g_i(q)^a + \ _i(q)) = \min f\ _g\ j\ F(\ _g) \qquad g \\
&= \min f\ _g\ j\ P(g_i(q)^a + \ _i(q) \qquad _g) \qquad g \\
&= \min f\ _g\ j\ P\ \frac{_i(q)}{} \qquad \frac{_g \quad g_i(q)^a}{} \qquad g \\
&= \min f\ _g\ j\ \frac{_g \quad g_i(q)^a}{} \qquad g:
\end{aligned} \tag{6}$$

(1) Computing safety bounds for inequality stochastic constraints with VaR on a manifold:



(a) (b)

(c)

Fig. 2: VaR safety bounds on a manifold for inequality constraints in terms of (a) safe probability, (b) constraints' mean, and (c) constraints' standard deviation.

Lemma 1. The safety bound BV_{g1} on a stochastic inequality constraint manifold via VaR is given by,

$$g_i(q)^a + g_i(q) \quad |\underbrace{\ ^1(\ ) \qquad}_{\text{safety bound}\ BV_{g1}}\{z \qquad + g_i(q)\}:$$

Proof. With Equation (6), we can have the following safety safety bound,

$$\begin{aligned}
&\frac{_g \quad g_i(q)^a}{} \\
&=) \quad \frac{_g \quad g_i(q)^a}{} \qquad ^1(\ ) \\
&=) \quad _g \qquad ^1(\ ) \ + g_i(q)^a + \qquad :
\end{aligned} \tag{7}$$

We can observe the following equation if we want to ensure total safety with probability ,

$$\begin{aligned}
0 &= \ _g \qquad ^1(\ ) \ + g_i(q)^a + \\
&=) \quad g_i(q)^a \qquad ^1(\ ) \qquad :
\end{aligned} \tag{8}$$

Therefore, the safety bound BV_{g1} via VaR that considered uncertain areas is given as follows:

$$g_i(q)^a + g_i(q) \quad |\underbrace{\ ^1(\ ) \qquad}_{\text{safety bound}\ BV_{g1}}\{z \qquad + g_i(q)\}: \tag{9}$$

□

Based on Lemma 1, we can have VaR safety bound for inequality constraints in terms of (a) safety probability, (b) constraints' mean and (c) constraints' standard deviation, as shown in Figure 2. It demonstrates the robustness of our safety bound under varying conditions of safety probability (a), mean of constraints (b), and standard deviation of constraints (c). Once we have the safety bounds, we can search for the policy on the tangent space of safety bounds, which helps ensure performance stability. We also provide CVaR safety bound for inequality constraints, for details, see the following Lemma 2.

(2) Computing safety bounds for inequality stochastic constraints with CVaR on a manifold:

Inspired by the reference [46], since the stochastic constraints subjects to a Gaussian distribution, we can have

$$= \frac{1}{\sqrt{2}} e^{\frac{1}{2}\frac{x - g_i(q)^a - u}{}^2}; \quad (10)$$

$$\text{CVaR} = E[X \mid X \quad _g] = g_i(q)^a + + \frac{^1( )}{1} \quad 0$$

$$=) \quad g_i(q)^a \quad \frac{^1( )}{1}; \quad (11)$$

Lemma 2. The safety bound BC_{g1} on an inequality stochastic constraint manifold via CVaR that considered uncertain areas is given as follows:

$$g_i(q)^a + g_i(q) \quad \underbrace{\frac{^1( )}{1} + g_i(q)}_{\text{safety bound } BC_{g1}}; \quad (12)$$

Proof. We have a random vector $X \quad 2 \quad R^{F+G}$, and $\text{CVaR}_g(X) = E(X \mid X > _g) = \frac{R_1}{g} x f(x)dx = (1 \quad F(_g))$, where $F(_g)$ is a CDF, $f(x)$ is a Probability Density Function (PDF), and $_g = \text{VaR}_g$, and $P(X > _g) = 1 \quad , x \; 2 \; R^{F+G}$.

As shown in the first part, $[(_g \quad) =] = \quad$, we have

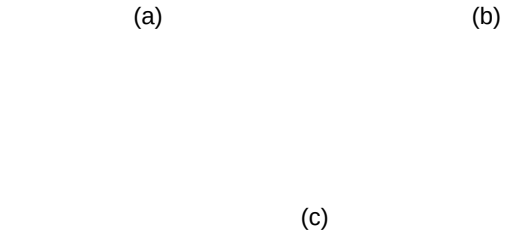
$$\begin{aligned}
&\int_{g}^{1} x f(x)dx = \int_{1}^{1} x f(x)dx \quad \int_{1}^{g} x f(x)dx \\
&= \int_{1}^{g} x f(x)dx \\
&= [(_g \quad) =] \int_{1}^{(_g \quad)=} z \; (z)dz \\
&= + \int_{1}^{(_g \quad)=} {}^0(z)dz \\
&= (1 \quad) + [(_g \quad) =] \\
&= (1 \quad) + \quad ^1( ) :
\end{aligned}$$

Thus, we have the safety bound,

$$\begin{aligned}
\text{CVaR}_g(X) &= \frac{(1 \quad) + \quad ^1( )}{1 \quad F(_g)} \\
&= \frac{(1 \quad) + \quad ^1( )}{1} \\
&= \quad + \frac{^1( )}{1}: \quad (13)
\end{aligned}$$

□

In accordance with Lemma 2, it becomes apparent that we are able to establish a safety bound predicated on CVaR for inequality constraints, and this relationship is explicated with respect to three fundamental components: (a) the safe probability, (b) the mean of the constraints, and (c) the standard deviation of the constraints. The representation of these relationships is provided in Figure 3.



(a) (b)

(c)

Fig. 3: CVaR safety bounds on a manifold for inequality constraints in terms of safe probability (a), constraints' mean (b), and constraints' standard deviation (c).

2) Safety Bounds for Equality Constraints with Stochastic Noises on a Manifold:

Stochastic equality constraints present a distinct challenge compared to stochastic inequality constraints due to their inherent two-sided constraint structure on a manifold, resulting in the existence of two separate safety bounds. The computation of probabilities for both sides of these equality constraints is a non-trivial task. To address this, we propose a transformation that allows us to reframe the problem of computing two equality safety bounds into two distinct problems centered around the computation of inequality safety bounds. Specifically, we aim for VaR and CVaR safety bounds based on the two inequality safety bounds obtained through this transformation.

Here, we rewrite VaR for equality constraints, $f_i(q)$ denotes equality constraints, as shown in Equations (14) and (15).

$$\begin{aligned}
F(_f) &= P(jf_i(q) + \; _i(q)j \quad _f) \quad (14) \\
&= P(\quad _f \quad f_i(q) + \; _i(q) \quad _f) \\
&= P(f_i(q) + \; _i(q) \quad _f) \quad P(f_i(q) + \; _i(q) \quad _f) \\
&= \quad \frac{_f \quad f_i(q)}{} \quad \frac{_f \quad f_i(q)}{} :
\end{aligned}$$

$$\begin{aligned}
&\text{VaR}_{}(jf_i(q) + \; _i(q)j) = \min \{ \; _f \mid F(_f) \quad \} \\
&= \min \{ \; _f \mid \frac{_f \quad f_i(q)}{} \quad \frac{_f \quad f_i(q)}{} \quad : \quad (15)
\end{aligned}$$

For one side safety bound via VaR, $P(_f^a) = P(f_i(q)^a + \; _i(q) > \; _f^a) < \frac{1}{2}$. For the other side safety bound via VaR, $P(_f^b) = P \quad f_i(q)^b + \; _i(q) \quad _f^b < \frac{1}{2}$, $_f^b$ is the other side's constrained limit set.

(1) Computing Safety Bounds for Equality Constraints with VaR on a Manifold:

Theorem 1. The first-side safety bound BV_{f1} on the equality stochastic constraint manifold via VaR that considers uncertain areas is given by,

$$f_i(q)^a + f_i(q) < \underbrace{\Phi^{-1}\left(\frac{1+\alpha}{2}\right)}_{\text{safety bound } BV_{f1}} + f_i(q):$$

The second-side safety bound BV_{f2} on the equality stochastic constraint manifold via VaR that considers uncertain areas is given by,

$$\underbrace{\Phi^{-1}\left(\frac{1-\alpha}{2}\right)}_{\text{safety bound } BV_{f2}} + f_i(q) < f_i(q)^b + f_i(q):$$

Proof. Based on Equations (14) and (15), we can prove the first side of safety bounds.

(A). The first side safety bound via VaR:

With the VaR definition on constraint manifolds (as shown in Definition 2) and Equation (5), we can have

$$\begin{aligned}
F(\alpha_f^a) &= P(f_i(q)^a + \alpha_i(q) > \alpha_f^a) < \frac{1-\alpha}{2} \\
&\Rightarrow 1 - P(f_i(q)^a + \alpha_i(q) \leq \alpha_f^a) < \frac{1-\alpha}{2} \\
&\Rightarrow \frac{1+\alpha}{2} < P\left(\frac{\alpha_i(q)}{\sigma_i(q)} \leq \frac{\alpha_f^a - f_i(q)^a}{\sigma_i(q)}\right) \\
&\Rightarrow \frac{1+\alpha}{2} < \Phi\left(\frac{\alpha_f^a - f_i(q)^a}{\sigma_i(q)}\right) \\
&\Rightarrow \Phi^{-1}\left(\frac{1+\alpha}{2}\right) < \frac{\alpha_f^a - f_i(q)^a}{\sigma_i(q)} \\
&\Rightarrow \Phi^{-1}\left(\frac{1+\alpha}{2}\right) + f_i(q)^a + \sigma_i(q) < \alpha_f^a: \qquad (16)
\end{aligned}$$

Since it's a equality constraint equation, we can have $\alpha_f^a = 0$, thus, the condition of uncertain areas of $f_i(q)^a$ is as follows,

$$\Phi^{-1}\left(\frac{1+\alpha}{2}\right) + f_i(q)^a + \sigma_i(q) < \alpha_f^a = 0 \qquad (17)$$

$$\Rightarrow f_i(q)^a < -\Phi^{-1}\left(\frac{1+\alpha}{2}\right): \qquad (18)$$

Therefore, the safety bound BV_{f1} via VaR that considered uncertain areas is given as follows:

$$f_i(q)^a + f_i(q) < \underbrace{\Phi^{-1}\left(\frac{1+\alpha}{2}\right)}_{\text{safety bound } BV_{f1}} + f_i(q): \qquad (19)$$

(B). The second side safety bound via VaR

Similar to the proof of the first side safety bound, we can have

$$\begin{aligned}
F(\alpha_f^b) &= P\left(f_i(q)^b + \alpha_i(q) \geq \alpha_f^b < \frac{1-\alpha}{2}\right) \\
&\Rightarrow P\left(\frac{\alpha_i(q)}{\sigma_i(q)} \leq \frac{\alpha_f^b - f_i(q)^b}{\sigma_i(q)}\right) < \frac{1-\alpha}{2} \\
&\Rightarrow \frac{\alpha_f^b - f_i(q)^b}{\sigma_i(q)} < \frac{1-\alpha}{2} \\
&\Rightarrow \alpha_f^b - f_i(q)^b < -\Phi^{-1}\left(\frac{1-\alpha}{2}\right) \\
&\Rightarrow f_i(q)^b - \Phi^{-1}\left(\frac{1-\alpha}{2}\right) < \alpha_f^b: \qquad (20)
\end{aligned}$$

The condition of uncertain areas of $f_i(q)^b$ is as follows,

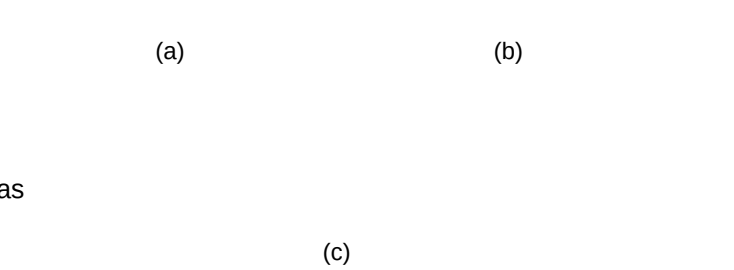
$$f_i(q)^b - \Phi^{-1}\left(\frac{1-\alpha}{2}\right) < \alpha_f^b = 0$$

$$\Rightarrow -\Phi^{-1}\left(\frac{1-\alpha}{2}\right) < f_i(q)^b: \qquad (21)$$

Thus, the safety bound BV_{f2} via VaR that considered uncertain areas is given as follows:

$$\underbrace{\Phi^{-1}\left(\frac{1-\alpha}{2}\right)}_{\text{safety bound } BV_{f2}} + f_i(q) < f_i(q)^b + f_i(q): \qquad (22)$$

□



(a) (b) (c)

Fig. 4: VaR safety bounds on a manifold for Equality Constraints in terms of safe probability (a), constraints' mean (b), and constraints' standard deviation (c).

Leveraging the aforementioned Theorem, we are able to establish VaR safety bounds specific to equality constraints. These bounds are characterized by three crucial parameters: the safe probability, the constraints' mean, and the constraints' standard deviation. A detailed representation of this relationship is depicted in Figure 4. Furthermore, we present the CVaR safety bounds pertinent to equality constraints. A detailed elaboration of these bounds is introduced in Theorem 2, providing a comprehensive analysis of safety bounds.

(2) Computing Safety Bounds for Equality Constraints with CVaR on a Manifold:

Theorem 2. For the first side, safety bound via CVaR:

$$\begin{aligned}
&\mathrm{CVaR}_{\alpha}[X \mid X \le \alpha_f] = f_i(q)^a + \frac{\phi(\Phi^{-1}(\frac{1+\alpha}{2}))}{\frac{1-\alpha}{2}} \\
&\Rightarrow f_i(q)^a + \frac{\phi(\Phi^{-1}(\frac{1+\alpha}{2}))}{\frac{1-\alpha}{2}} - X = 0 \\
&\Rightarrow f_i(q)^a \le \frac{\phi(\Phi^{-1}(\frac{1+\alpha}{2}))}{\frac{1-\alpha}{2}}. \qquad (23)
\end{aligned}$$

The first side-safety bound BC_{f1} on equality stochastic constraint Manifolds via CVaR that considered uncertain areas is given as follows:

$$f_i(q)^a + f_i(q) \le \underbrace{\frac{\phi(\Phi^{-1}(\frac{1+\alpha}{2}))}{\frac{1-\alpha}{2}}}_{\text{safety bound } BC_{f1}} + f_i(q). \qquad (24)$$

For the second side safety bound via CVaR:

$$\begin{aligned}
&\mathrm{CVaR}_{\alpha}[X \mid X \ge \alpha_f] \\
&= f_i(q)^b + \frac{\phi(\Phi^{-1}(\frac{1-\alpha}{2}))}{\frac{1-\alpha}{2}} \\
&\Rightarrow f_i(q)^b + \frac{\phi(\Phi^{-1}(\frac{1-\alpha}{2}))}{\frac{1-\alpha}{2}} - X = 0 \\
&\Rightarrow f_i(q)^b \ge -\frac{\phi(\Phi^{-1}(\frac{1-\alpha}{2}))}{\frac{1-\alpha}{2}}. \qquad (25)
\end{aligned}$$

The second side-safety bound BC_{f2} on equality stochastic constraint manifolds via CVaR that considered uncertain areas is given as follows:

$$f_i(q)^b + f_i(q) \ge \underbrace{-\frac{\phi(\Phi^{-1}(\frac{1-\alpha}{2}))}{\frac{1-\alpha}{2}}}_{\text{safety bound } BC_{f2}} + f_i(q). \qquad (26)$$

Proof. On the basis of Lemma 2 and Theorem 1, for a random vector X within equality constraints on a manifold, CVaR is denoted as $\mathrm{CVaR}_{\alpha}(X) = E(X \mid X > \alpha_f) = \int_{\alpha_f}^{1} x f(x)dx/(1-F(\alpha_f))$, and $\alpha_f = \mathrm{VaR}_{\alpha}$, and $P(X > \alpha_f) = (1-\alpha)/2$ [46].

Since $X \sim N(\mu, \sigma^2)$, $Z = (X-\mu)/\sigma \sim N(0,1)$, the standard normal distribution function specifies $\Phi(z) = P(Z \le z)$ and $1-\Phi(z) = P(Z > z)$.

(A). The first side safety bound via CVaR:

$$\begin{aligned}
\int_{\alpha_f}^{1} x f(x)dx &= \int_{-\infty}^{1} x f(x)dx - \int_{-\infty}^{\alpha_f} x f(x)dx \\
&= \int_{-\infty}^{\alpha_f} x f(x)dx.
\end{aligned}$$

Changing variables with $x = \mu + z\sigma$ and $dx = \sigma dz$. Based on the change of variables, noting that $\phi(z) = \sigma f(\mu + z\sigma)$, is a PDF which subjects to a Gaussian distribution, we have

$$\begin{aligned}
\int_{\alpha_f}^{1} x f(x)dx &= \int_{-\infty}^{\Phi(\alpha_f)} (\mu + z\sigma)\phi(z)dz \\
&= \int_{-\infty}^{1} \mu\phi(z)dz + \int_{-\infty}^{\Phi(\alpha_f)} z\sigma\phi(z)dz \\
&= \mu\alpha_f = \mu + \int_{-\infty}^{\Phi(\alpha_f)} z\sigma\phi(z)dz.
\end{aligned}$$

As shown in the above Equations, $\alpha_f = \mu = \mu + (1-\alpha)/2 = (1+\alpha)/2$, and the key here is the observation that the standard normal density function satisfies $\phi(z) = \phi'(z)$.

$$\begin{aligned}
&\int_{\alpha_f}^{1} x f(x)dx \\
&= \mu\alpha_f = \int_{-\infty}^{\Phi(\alpha_f)} z\sigma\phi(z)dz \\
&= \frac{1+\alpha}{2} + \int_{-\infty}^{\Phi(\alpha_f)} \sigma\phi'(z)dz \\
&= (\frac{1}{2}) + \alpha_f = \\
&= (\frac{1}{2}) + \Phi^{-1}(\frac{1+\alpha}{2}).
\end{aligned}$$

With the above Equations the definition of CVaR (as shown in Definition 2), we can have the following safety bounds,

$$\begin{aligned}
\mathrm{CVaR}_{\alpha_f}(X) &= \frac{(\frac{1}{2}) + \Phi^{-1}(\frac{1+\alpha}{2})}{1-\alpha_f} \\
&= \frac{(\frac{1}{2}) + \Phi^{-1}(\frac{1+\alpha}{2})}{\frac{1-\alpha}{2}} \\
&= \mu + \frac{\phi(\Phi^{-1}(\frac{1+\alpha}{2}))}{\frac{1-\alpha}{2}}.
\end{aligned}$$

(B). The second side safety bound via CVaR:

With proof of the first side CVaR safety bound, we can easily have the second side CVaR safety bound, which is shown as follows:

$$\begin{aligned}
\int_{-\infty}^{\alpha_f} x f(x)dx &= E[z < (\alpha_f - \mu)/\sigma] \\
&= \int_{-\infty}^{\Phi(\alpha_f)} (\mu + z\sigma)\phi(z)dz \\
&= \int_{-\infty}^{1} \mu\phi(z)dz + \int_{-\infty}^{\Phi(\alpha_f)} z\sigma\phi(z)dz \\
&= \mu\alpha_f = \mu + \int_{-\infty}^{\Phi(\alpha_f)} z\sigma\phi(z)dz \\
&= (\frac{1}{2})\mu - \alpha_f = \\
&= (\frac{1}{2})\mu - \Phi^{-1}(\frac{1-\alpha}{2}),
\end{aligned}$$

(a) (b)

(c)

Fig. 5: CVaR safety bounds on a manifold for equality constraints in terms of safe probability (a), constraints' mean (b), and constraints' standard deviation (c).

$$\begin{aligned}\mathrm{CVaR}_{\frac{b}{f}}(X) &= \frac{(\frac{1}{2}) \quad [\quad ^{1}(\frac{1}{2})]}{F(\frac{b}{f})} \\ &= \frac{(\frac{1}{2}) \quad [\quad ^{1}(\frac{1}{2})]}{\frac{1}{2}} \\ &= \frac{[\quad ^{1}(\frac{1}{2})]}{\frac{1}{2}}.\end{aligned}$$

□

Thus, we can have CVaR safety bound for equality constraints in terms of safe probability, constraints' mean and constraints' standard deviation, as shown in Figure 5.

In the following section, we will introduce a practical algorithm aimed at effectively addressing and managing these stochastic constraints.

C. Practical Algorithm

Building upon the theoretical foundations outlined earlier, we have devised a practical algorithm, as detailed in Algorithm 1, designed to ensure safety in complex environments characterized by stochastic constraints. The algorithm unfolds in three key steps. Initially, stochastic constraints are formulated within the framework of a manifold. Subsequently, safety bounds are computed through the utilization of VaR and CVaR metrics. Finally, the algorithm projects the safe operational space onto the manifold and orchestrates the search for a secure policy that adheres to the stochastic constraint manifold.

Notably, to simplify computation, we leverage Gaussian noises to represent stochastic constraints $c_i(q)$, which can be easily learned via Gaussian processes [47]. As shown in Equation (27), f represents the stochastic constraints that equal stochastic constraints $c_i(q)$ plus deterministic constraints $c_i(q)$, and we need to compute its safety bounds given $c_i(q)$ and $_i(q)$.

$$f = c_i(q) + \quad _i(q) \quad _i(q) \quad N \quad 0; \ ^2_n : \tag{27}$$

The stochastic constraints are assumed as independent, and the constraints are subject to independent distributions.

Algorithm 1 Searching a RObust Safe Policy on a Stochastic COnstraint Manifold (ROSCOM)

1: Input stochastic constraints $c_i(q) + \ _i$.
2: Initialise a policy.
3: **for** $k = 0; 1; : : : ; T$ **do**
4: Compute the stochastic constraints $c_i(q) + \ _i$ at each step.
5: Compute the VaR and CVaR bounds for high dimensional inequality and equality constraints.
6: Project the safety bounds into the manifold space $c(q; \ )$.
7: Compute the Jacobian matrix $J_i(q; \ _s) = \frac{@}{@q}\overline{J_i}(q; \ _s)^>; \frac{@}{@}\overline{J_i}(q; \ _s)^> \ 2 \ R^{(n+N) \ N}$.
8: Compute the null space $N(q; \ _s)$.
9: Sampling trajectories on a tangent space of a constrained manifold with RL algorithms.
10: Compute action $a_k = (\ q_k)$ based on ATACOM.
11: Deploy action a_k in the environment and provide the reward and observation to the RL algorithm.
12: **end for**

V. EXPERIMENTS

In this section, we conduct a series of experiments with the primary objective of evaluating the effectiveness of our proposed method. Furthermore, we undertake a comparative analysis to assess the performance of our approach in relation to existing SOTA safe RL baselines.

Specifically, we first compare our method with ATACOM [1] on circular motion tasks. Furthermore, to comprehensively evaluate the effectiveness of our methods, we compare our method with ATACOM and traditional CMDP algorithms on challenging air-hockey tasks. Generally, ATACOM requires constraint information, which is leveraged to guarantee safety most rigorously. For example, the constraint information is human posture in a human-robot interaction environment. However, traditional CMDP algorithms do not require constraint information, and they ensure safety by try-and-error learning, which could be helpful in uncritical safety environments. The representative SOTA algorithms of traditional CMDP settings are CPO [2], PCPO [3], and so on. The environment settings are provided in Appendix VII-A and details of implementation are introduced in Appendix VII-B

A. Circular Motion Tasks

In the circular motion task, a robot must run on the blue circle as shown in Figure 6. If the robot runs away from the circle, it will incur a cost. However, the constraint is stochastic. The blue circle denotes the original constraint, the blue points denote the stochastic constraints, and the two green circles denote the safety bounds via VaR, the two red circles represent the safety bounds via CVaR for the new constraints.

1) Compare with ATACOM: The experimental results are presented in Figures 7 (a) and (b), where c_{avg} denotes the average cost of each trajectory, denotes the reward performance, ROSCOM-PPO and ATACOM-PPO [1] denote

(a) Safety Performance (b) Reward Performance

Fig. 7: Compare our method with ATACOM [1] on a circular motion task regarding safety (lower values signify superior performance) and reward (higher values indicate better performance) performance.

(a) Safety Performance (b) Reward Performance

Fig. 8: Compare our method with traditional CMDP methods, CPO [2] and PCPO [3], on a circular motion task regarding safety and reward performance.

our ROSCOM with PPO methods [48], and ATACOM with PPO methods. Although the ATACOM method presents remarkable performance for ensuring robot learning safety, a comparative analysis of experimental outcomes suggests that our proposed methodology outperforms ATACOM. More precisely, the proposed method demonstrates superior reward and enhanced safety performance while taking into consideration the stochastic constraints.

2) Compare with Traditional CMDP Algorithms: In Figure 8, we present the results of comparison experiments focused on tasks involving circular motion. Specifically, as depicted in Figure 8(a), our proposed method not only maintains rigorous safety standards but also significantly outperforms traditional safe RL baselines, including representative baselines CPO [2] and PCPO [3]. These baselines fail to consistently ensure safety during the learning process. Further, Figure 8(b) illustrates that our method achieves performance comparable to that of the safe RL baselines. These experimental results demonstrate that our approach consistently surpasses traditional safe RL baselines in terms of the balance between safety and reward performance, indicating its efficacy and reliability in safe RL applications.

Fig. 6: Safety bounds on a stochastic constraint manifold. The green and red circles denote the safety bounds via VaR and CVaR. In stochastic constraints with Gaussian parameters, the mean is 0, the standard deviation is 0.05, and the safety probability is 99%.

B. Air-Hockey Tasks

In the Air-Hockey task, as shown in Figure 9, the robot needs to learn to manipulate its hand and hit a lightweight plastic puck to score goals. However, due to noises, e.g., sensor measurement errors, the table height estimation by sensors can vary randomly. To prevent the robot from colliding with the table and causing damage to the robot and the table, we need to compute safety bounds for robot learning, so that the robot can complete the task safely, even with stochastic constraints.

1) Compare with ATACOM:As depicted in Figure 10, we compare our method with ATACOM, the experimental results demonstrate that our method is better than ATACOM in terms of safety and reward performance, e.g., the average cost value of trajectories is lower than ATACOM (Figure 10 (a)), and reward value is higher than ATACOM (Figure 10 (b)). The results indicate that our method can be more rigorous than ATACOM to ensure safety for critical applications.

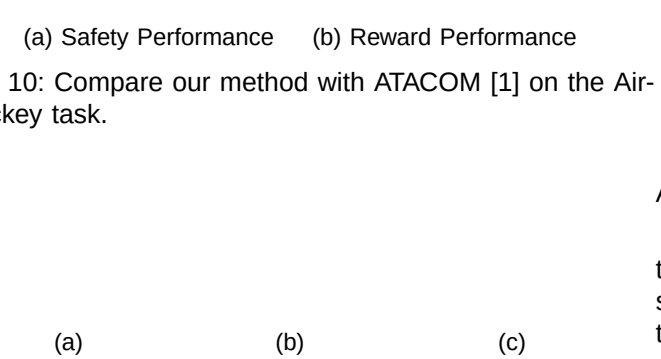
Fig. 9: An Air-Hockey task environment.

2) Compare with Traditional CMDP Algorithms: In order to provide a comprehensive evaluation of the efficacy of our proposed method, we extend our assessment by conducting comparative experiments. Specifically, we compare our method with SOTA representative safe RL baselines, including CPO [2] and PCPO [3].

The evaluation results, as depicted in Figure 11, reveal the remarkable superiority of our method in comparison to SOTA baselines regarding safety and reward performance. Firstly, our method exhibits significantly improved safety performance, as evidenced by the average cost values of each trajectory, as illustrated in Figure 11 (a). Secondly, in terms of reward performance, our approach outperforms the strong baselines, as demonstrated by Figures 11 (b) and (c), with Figure 11 (c) providing an enlarged view of Figure 11 (b) in terms of SOTA baselines' reward performance. Our methodology's capability to achieve superior outcomes is clearly demonstrated, showing its potential for robust and safe RL real-world applications.

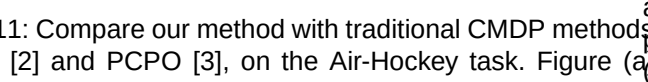
The distinct superiority of our method over traditional algorithms can be attributed to several key elements integrated within our approach. Firstly, the constrained space is leveraged to construct a constrained manifold, a feature that proves to be instrumental in facilitating efficient high-dimensional learning. Secondly, the tangent space of the constrained manifold is explicitly established for policy searching. It provides a rigorous safety assurance mechanism within the policy search process. By operating within the tangent space, the policy search is inherently aligned with safety protocols, ensuring that each iteration adheres to safety constraints.

These key points collectively contribute to the significant performance improvement of our method, positioning it as



(a) Safety Performance (b) Reward Performance

Fig. 10: Compare our method with ATACOM [1] on the Air-Hockey task.



(a) (b) (c)

Fig. 11: Compare our method with traditional CMDP methods, CPO [2] and PCPO [3], on the Air-Hockey task. Figure (a) shows the safety performance, Figures (b) and (c) show the reward performance.

a notably better solution than traditional algorithms, and promising enhanced safety and efficiency in complex, high-dimensional learning environments.

Influence of Safety Bounds on Reward Performance: (1) In our experiments, the safety bound significantly reduces learning oscillations during policy search, enhancing overall policy performance. For example, in the circular motion task, the state space's stochastic noise can lead to policy oscillations in the baseline algorithm such as ATACOM, subsequently degrading policy performance. Similar effects are observed in air hockey tasks. Unlike these methods, our approach establishes a stable safety bound that accommodates the uncertainty of the state space. This allows for more effective policy exploration than baseline methods, thereby improving safety and reward performance. (2) Compared to traditional safe RL methods, our approach is developed on constrained manifolds. This design facilitates safe exploration within the tangent space and guides the policy toward a higher-quality outcome. Consequently, our method achieves superior safety and reward performance relative to traditional approaches.

VI. Conclusion

In this paper, we investigated the problem of robust safe RL on stochastic constraint manifolds. Prior safe RL research has focused on deterministic constraints, whereas in this paper we considered stochastic constraints. First, we formulated the robust safe RL problem by considering the stochastic constraints. Second, we ensured safety by leveraging the risk measurement methods, e.g., VaR and CVaR. Safety bounds are computed on the stochastic constraint manifold to help search for a safe policy. Finally, we evaluated our method on the circular motion and Air-Hockey tasks. The experiment results demonstrate that our algorithm can achieve remarkable performance in terms of learning safety, and show better reward performance than the SOTA baselines. In the Future, we plan to investigate the model's efficiency and try to deploy our method in real-world robot control.

Acknowledgments

We thank Dr. Davide Tateo for his helpful discussions.

Appendix

VII. Details of Experiments

A. Environment Settings

Circle Motion Environments. In this task, the robot needs to run as fast as possible to reach the goal position while satisfying safety constraints. As shown in the following equations, $R(X_t; Y_t)$ denotes the reward value when the robot is at position $(X_t; Y_t)$, X_t denotes the robot X-axis direction position and Y_t denotes its Y-axis direction position, similarly, $(X_g; Y_g)$ denotes the goal position, denotes the stochastic constraints, and $C_{inequality}$ denotes the inequality constraints. other settings are similar to ATACOM [1].

$$R(X_t; Y_t) = \exp(\quad \sqrt{(X_t \quad X_g)^2 + (Y_t \quad Y_g)^2}) \tag{28}$$

$$C(X_t; Y_t) = j \sqrt{X_t^2 + Y_t^2} \quad (1 + \ )j + C_{inequality} \tag{29}$$

$$C_{inequality} = \begin{cases} jY_t + 0{:}5 + \ j; & Y_t < \quad (0{:}5 + \ ); \\ 0; & \text{Others:} \end{cases} \tag{30}$$

Air-Hockey Environments.
Other reward and constraint settings of the task are same to ATACOM [1], except for Equation (31), Z_{table} denotes the table height, Z_t denotes the robot Z-axis direction position, puck height is 0.0149m. We provide more experiments to evaluate the effectiveness of our method regarding the different safety limits, as shown in Figure 12, the experiment results indicate that our method can perform remarkably better than SOTA safe RL baselines regarding reward and safety performance, and confirm the effectiveness of our method again. The implementation of CPO and PCPO from a repository of safe RL baselines[1] is used to carry out the comparison experiments.
Our method outperforms the SOTA safe RL baselines because the safety constraints are learned and projected on a Manifold space, which means our algorithm can receive the safety and task environment information in advance, which could help our method better search policy while satisfying safety constraints.

$$C_{inequality} = \begin{cases} jZ_t \quad (Z_{table} + \ )j; & Z_t < (Z_{table} + \ ); \\ 0; & \text{Others:} \end{cases} \tag{31}$$

(a) (b) (c)

Fig. 12: Compare our method with traditional CMDP methods, CPO [2] and PCPO [3], on the Air-Hockey task with different safety bounds. Figure (a) shows the safety performance, Figures (b) and (c) show the reward performance.

Parameters	value	Parameters	value
actor lr	3e-4	critic lr	3e-4
batch size	64	gamma	0.99
horizon	500	epoch	50
eps ppo	0.1	noise mean	0
	99%	noise std	0.5
network	MLP	regular	relu & linear
NN features	[32, 32]		

TABLE I: ROSCOM and ATACOM hyparameters used in Circle Motion experiments.

Parameters	value	Parameters	value
actor lr	3e-4	critic lr	3e-4
batch size	64	gamma	0.99
horizon	3000	epoch	200
eps ppo	0.1	noise mean	0
	99%	noise std	0.5
network	MLP	regular	relu & linear
NN features	[64, 64]		

TABLE II: Algorithms' hyparameters of ROSCOM, ATACOM, CPO, PCPO, used in Air-Hockey experiments. Safety bound used for CPO and PCPO in Figures 8 and 12 is 25, and in Figure 12 is 5.

B. Details of Implementing Experiments

The parameters used in our experiments are provided in Tables I and II.

REFERENCES

[1] P. Liu, D. Tateo, H. B. Ammar, and J. Peters, "Robot reinforcement learning on the constraint manifold," in Conference on Robot Learning PMLR, 2022, pp. 1357–1366.

[2] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in International conference on machine learning PMLR, 2017, pp. 22–31.

[3] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," in International Conference on Learning Representations, 2020.

[4] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, Y. Yang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," arXiv preprint arXiv:2205.10330, 2022.

[5] J. Yang, J. Ni, M. Xi, J. Wen, and Y. Li, "Intelligent path planning of underwater robot based on reinforcement learning," IEEE Transactions on Automation Science and Engineering, 2022.

[6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., "Mastering the game of go with deep neural networks and tree search," nature, vol. 529, no. 7587, pp. 484–489, 2016.

[7] N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk et al., "Optimizing debt collections using constrained reinforcement learning," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 75–84.

[8] S. Gu, J. G. Kuba, Y. Chen, Y. Du, L. Yang, A. Knoll, and Y. Yang, "Safe multi-agent reinforcement learning for multi-robot control," Artificial Intelligence, vol. 319, p. 103905, 2023.

[9] S. Gu, D. Huang, M. Wen, G. Chen, and A. Knoll, "Safe multi-agent learning with soft constrained policy optimization in real robot control," IEEE Transactions on Industrial Informatics, 2024.

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

[11] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in International Conference on Learning Representations, 2018.

[12] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," science, vol. 290, no. 5500, pp. 2319–2323, 2000.

[13] E. S. Gastal and M. M. Oliveira, "Adaptive manifolds for real-time high-dimensional filtering," ACM Transactions on Graphics (TOG), vol. 31, no. 4, pp. 1–13, 2012.

[14] A. J. Izenman, "Introduction to manifold learning," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 4, no. 5, pp. 439–446, 2012.

[15] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a cvar optimization approach," Advances in neural information processing systems, vol. 28, 2015.

[16] C. Ying, X. Zhou, D. Yan, and J. Zhu, "Towards safe reinforcement learning via constraining conditional value at risk," in IJCAI 2021 Workshop on Adversarial Machine Learning, 2021.

[17] S. Gu, A. Kshirsagar, Y. Du, G. Chen, J. Peters, and A. Knoll, "A human-centered safe robot reinforcement learning framework with interactive behaviors," Frontiers in Neurorobotics, vol. 17, 2023.

[18] Z. Liu, Z. Guo, Z. Cen, H. Zhang, J. Tan, B. Li, and D. Zhao, "On the robustness of safe reinforcement learning under observational perturbations," ICLR, 2023.

[19] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration in finite markov decision processes with gaussian processes," Advances in Neural Information Processing Systems, vol. 29, 2016.

[20] F. Berkenkamp and A. P. Schoellig, "Safe and robust learning control with gaussian processes," in 2015 European Control Conference (ECC) IEEE, 2015, pp. 2496–2501.

[21] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with gaussian processes," in international conference on machine learning PMLR, 2015, pp. 997–1005.

[22] A. Wachi, Y. Sui, Y. Yue, and M. Ono, "Safe exploration and optimization of constrained mdps using gaussian processes," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[23] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," Advances in neural information processing systems, vol. 31, 2018.

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

[25] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in 2018 IEEE conference on decision and control (CDC) IEEE, 2018, pp. 6059–6066.

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

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

[28] N. Fulton and A. Platzer, "Safe reinforcement learning via formal methods: Toward safe control through proof and learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[29] X. Liang, K. Shu, K. Lee, and P. Abbeel, "Reward uncertainty for exploration in preference-based reinforcement learning," International Conference on Learning Representations

[30] K. Li, A. Gupta, A. Reddy, V. H. Pong, A. Zhou, J. Yu, and S. Levine, "Mural: Meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning," in International conference on machine learning PMLR, 2021, pp. 6346–6356.

[31] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," Advances in neural information processing systems, vol. 31, 2018.

[1] https://github.com/PKU-Alignment/Safe-Policy-Optimization.git