Safe Exploration Method for Reinforcement Learning under Existence of Disturbance

Yoshihiro Okawa (✉)[1][0000−0001−5095−4927], Tomotake Sasaki[1][0000−0002−3376−2779], Hitoshi Yanami[1], and Toru Namerikawa[2][0000−0001−9907−4234]

[1] Artificial Intelligence Laboratory, Fujitsu Limited, Kawasaki, Japan
{okawa.y,tomotake.sasaki,yanami}@fujitsu.com
[2] Department of System Design Engineering, Keio University, Yokohama, Japan
namerikawa@keio.jp

Abstract. Recent rapid developments in reinforcement learning algorithms have been giving us novel possibilities in many fields. However, due to their exploring property, we have to take the risk into consideration when we apply those algorithms to safety-critical problems especially in real environments. In this study, we deal with a safe exploration problem in reinforcement learning under the existence of disturbance. We define the safety during learning as satisfaction of the constraint conditions explicitly defined in terms of the state and propose a safe exploration method that uses partial prior knowledge of a controlled object and disturbance. The proposed method assures the satisfaction of the explicit state constraints with a pre-specified probability even if the controlled object is exposed to a stochastic disturbance following a normal distribution. As theoretical results, we introduce sufficient conditions to construct conservative inputs not containing an exploring aspect used in the proposed method and prove that the safety in the above explained sense is guaranteed with the proposed method. Furthermore, we illustrate the validity and effectiveness of the proposed method through numerical simulations of an inverted pendulum and a four-bar parallel link robot manipulator.

Keywords: Reinforcement learning · Safe exploration · Chance constraint.

1 Introduction

Guaranteeing safety and performance during learning is one of the critical issues to implement reinforcement learning (RL) in real environments [12,14]. To address this issue, RL algorithms and related methods dealing with safety have been studied in recent years and some of them are called “safe reinforcement learning” [10]. For example, Biyik et al. [4] proposed a safe exploration algorithm for a deterministic Markov decision process (MDP) to be used in RL. They guaranteed to prevent states from being unrecoverable by leveraging the Lipschitz continuity of its unknown transition dynamics. In addition, Ge et al. [11] proposed a modified Q-learning method for a constrained MDP solved with the Lagrange multiplier method so that their algorithm seeks for the optimal solution ensuring that the safety premise is satisfied. Several methods use prior knowledge of the controlled object (i.e., environment) for guaranteeing the safety [3,17]. However, few studies evaluated their safety quantitatively from a viewpoint of satisfying state constraints at each timestep that are defined explicitly in the problems. Evaluating safety from this viewpoint is often useful when we have constraints on a physical system and need to estimate the risk caused by violating those constraints beforehand.

Recently, Okawa et al. [19] proposed a safe exploration method that is applicable to existing RL algorithms. They quantitatively evaluated the above-mentioned safety in accordance with probabilities of satisfying the explicit state constraints. In particular, they theoretically showed that their proposed method assures the satisfaction of the state constraints with a pre-specified probability by using partial prior knowledge of the controlled object. However, they did not consider the existence of external disturbance, which is an important factor when we consider safety. Such disturbance sometimes makes the state violate the constraints even if the inputs (i.e., actions) used in exploration are designed to satisfy those constraints. Furthermore, they made a strong assumption regarding the controlled objects such that the state remains within the area satisfying the constraints if the input is set to be zero as a conservative input, i.e., an input that does not contain an exploring aspect.

In this study, we extend Okawa et al.'s work [19] and tackle the safe exploration problem in RL under the existence of disturbance[1]. Our main contributions are the following.

- We propose a novel safe exploration method for RL that uses partial prior knowledge of both the controlled object and disturbance.
- We introduce sufficient conditions to construct conservative inputs not containing an exploring aspect used in the proposed method. Moreover, we theoretically prove that our proposed method assures the satisfaction of explicit state constraints with a pre-specified probability under the existence of disturbance that follows a normal distribution.

We also demonstrate the validity and effectiveness of the proposed method with the simulated inverted pendulum provided in OpenAI Gym [6] and a four-bar parallel link robot manipulator [18] with additional disturbances.

The rest of this paper is organized as follows. In Section 2, we introduce the problem formulation of this study. In Section 3, we describe our safe exploration method. Subsequently, theoretical results about the proposed method are shown in Section 4. We illustrate the results of simulation evaluation in Section 5. We discuss the limitations of the proposed method in Section 6, and finally, we conclude this paper in Section 7.

[1] Further comparison with other related works is given in Appendix A (electronic supplementary material).

2 Problem formulation

We consider an input-affine discrete-time nonlinear dynamic system (environment) expressed by the following state transition equation:

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k) + \boldsymbol{G}(\boldsymbol{x}_k)\boldsymbol{u}_k + \boldsymbol{w}_k, \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^n$, $\boldsymbol{u}_k \in \mathbb{R}^m$, and $\boldsymbol{w}_k \in \mathbb{R}^n$ stand for the state, input (action) and disturbance at timestep k, respectively, and $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are unknown nonlinear functions. We suppose that the state $\boldsymbol{x}_k$ is directly observable. An immediate cost $c_{k+1} \geq 0$ is given depending on the state, input and disturbance at each timestep k:

$$c_{k+1} = c(\boldsymbol{x}_k,\ \boldsymbol{u}_k,\ \boldsymbol{w}_k), \tag{2}$$

where the immediate cost function $c : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to [0, \infty)$ is unknown while c_{k+1} is supposed to be directly observable. We consider the situation where the constraints that the state is desired to satisfy from the viewpoint of safety are explicitly given by the following linear inequalities:

$$\boldsymbol{H}\boldsymbol{x} \preceq \boldsymbol{d}, \tag{3}$$

where $\boldsymbol{d} = [d_1, \ldots, d_{n_c}]^\top \in \mathbb{R}^{n_c}$, $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{n_c}]^\top \in \mathbb{R}^{n_c \times n}$, n_c is the number of constraints and $\preceq$ means that the standard inequality $\leq$ on $\mathbb{R}$ holds for all elements. In addition, we define $\mathcal{X}_s \subset \mathbb{R}^n$ as the set of safe states, that is,

$$\mathcal{X}_s := \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{H}\boldsymbol{x} \preceq \boldsymbol{d}\}. \tag{4}$$

Initial state $\boldsymbol{x}_0$ is assumed to satisfy $\boldsymbol{x}_0 \in \mathcal{X}_s$ for simplicity.

The primal goal of reinforcement learning is to acquire a policy (control law) that minimizes or maximizes an evaluation function with respect to the immediate cost or reward, using them as cues in its trial-and-error process [20]. In this study, we consider the standard discounted cumulative cost as the evaluation function to be minimized:

$$J = \sum_{k=0}^{T} \gamma^k c_{k+1}. \tag{5}$$

Here, γ is a discount factor $(0 < \gamma \leq 1)$ and T is the terminal time.

Besides (5) for the cost evaluation, we define the safety in this study as satisfaction of the state constraints and evaluate its guarantee quantitatively. In detail, we consider the following chance constraint with respect to the satisfaction of the explicit state constraints (3) at each timestep k:

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \geq \eta, \tag{6}$$

where $\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\}(= \Pr\{\boldsymbol{x}_k \in \mathcal{X}_s\})$ denotes the probability that $\boldsymbol{x}_k$ satisfies the constraints (3).

The objective of the proposed safe exploration method is to make the chance constraint (6) satisfied at every timestep $k = 1, 2, \dots, T$ for a pre-specified η, where $0.5 < \eta < 1$ in this study.

Figure 1 shows the overall picture of the reinforcement learning problem in this study. The controller (agent) depicted as the largest red box generates an input (action) $\boldsymbol{u}_k$ according to a base policy with the proposed safe exploration method and apply it to the controlled object (environment) depicted as the green box, which is a discrete-time nonlinear dynamic system exposed to a disturbance $\boldsymbol{w}_k$. According to an RL algorithm, the base policy is updated based on the state $\boldsymbol{x}_{k+1}$ and immediate cost c_{k+1} observed from the controlled object. In addition to updating the base policy to minimize the evaluation function, the chance constraint should be satisfied at every timestep $k = 1, 2, \dots, T$. The proposed method is described in detail in Sections 3 and 4.

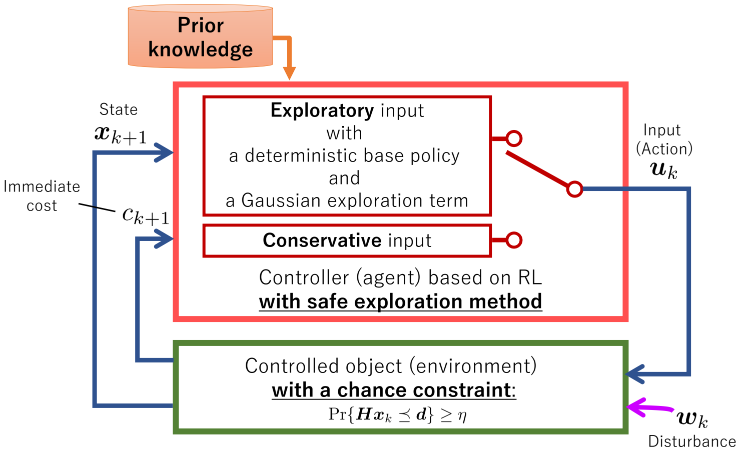
Prior knowledge
State
$\boldsymbol{x}_{k+1}$
Immediate cost
c_{k+1}
Exploratory input with a deterministic base policy and a Gaussian exploration term
Conservative input
Controller (agent) based on RL with safe exploration method
Input (Action)
$\boldsymbol{u}_k$
Controlled object (environment) with a chance constraint:
$\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \geq \eta$
$\boldsymbol{w}_k$
Disturbance

Fig. 1. Overview of controlled object (environment) under existence of disturbance and controller (agent) based on an RL algorithm with the proposed safe exploration method. The controller updates its base policy through an RL algorithm, while the proposed safe exploration method makes the chance constraint of controlled object satisfied by adjusting its exploration process online.

As the base policy, we consider a nonlinear deterministic feedback control law

$$\begin{aligned} \boldsymbol{\mu}(\,\cdot\,;\boldsymbol{\theta}) : \mathbb{R}^n &\to \mathbb{R}^m \\ \boldsymbol{x} &\mapsto \boldsymbol{\mu}(\boldsymbol{x};\boldsymbol{\theta}), \end{aligned} \tag{7}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{N_{\boldsymbol{\theta}}}$ is an adjustable parameter to be updated by an RL algorithm. When we allow exploration, we generate an input $\boldsymbol{u}_k$ by the following equation:

$$\boldsymbol{u}_k = \boldsymbol{\mu}(\boldsymbol{x}_k;\boldsymbol{\theta}_k) + \boldsymbol{\varepsilon}_k, \tag{8}$$

where $\boldsymbol{\varepsilon}_k \in \mathbb{R}^m$ is a stochastic exploration term that follows an m-dimensional normal distribution (Gaussian probability density function) with mean $\mathbf{0} \in \mathbb{R}^m$ and variance-covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{m \times m}$, denoted as $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$. In this case, as a consequence of the definition, $\boldsymbol{u}_k$ follows a normal distribution $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}_k; \boldsymbol{\theta}_k), \boldsymbol{\Sigma}_k)$.

We make the following four assumptions about the controlled object and the disturbance. The proposed method uses these prior knowledge to generate inputs, and the theoretical guarantee of satisfying the chance constraint is proven by using these assumptions.

Assumption 1 *Matrices $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ in the following linear approximation model of the nonlinear dynamics (1) are known:*

$$\boldsymbol{x}_{k+1} \simeq \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k. \tag{9}$$

The next assumption is about the disturbance.

Assumption 2 *The disturbance $\boldsymbol{w}_k$ stochastically occurs according to an n-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, where $\boldsymbol{\mu}_w \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_w \in \mathbb{R}^{n \times n}$ are the mean and the variance-covariance matrix, respectively. The mean $\boldsymbol{\mu}_w$ and variance-covariance matrix $\boldsymbol{\Sigma}_w$ are known, and the disturbance $\boldsymbol{w}_k$ and exploration term $\boldsymbol{\varepsilon}_k$ are uncorrelated at each timestep k.*

We define the difference $\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R}^n$ between the nonlinear system (1) and the linear approximation model (9) (i.e., approximation error) as below:

$$\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u}) := \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x})\boldsymbol{u} - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}). \tag{10}$$

We make the following assumption on this approximation error.

Assumption 3 *Regarding the approximation error $\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u})$ defined by (10), $\bar{\delta}_j < \infty$, $\bar{\Delta}_j < \infty$, $j = 1, \ldots, n_c$ that satisfy the following inequalities are known:*

$$\bar{\delta}_j \geq \sup_{\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{u} \in \mathbb{R}^m} |\boldsymbol{h}_j^\top \boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u})|, \quad j = 1, 2, \ldots, n_c, \tag{11}$$

$$\bar{\Delta}_j \geq \sup_{\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{u} \in \mathbb{R}^m} |\boldsymbol{h}_j^\top \left(\boldsymbol{A}^{\tau-1} + \boldsymbol{A}^{\tau-2} + \cdots + \boldsymbol{I}\right) \boldsymbol{e}(\boldsymbol{x}, \boldsymbol{u})|, \ j = 1, 2, \ldots, n_c, \tag{12}$$

where τ is a positive integer.

The following assumption about the linear approximation model and the constraints is also made.

Assumption 4 *The following condition holds for $\boldsymbol{B}$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{n_c}]^\top$:*

$$\boldsymbol{h}_j^\top \boldsymbol{B} \neq \mathbf{0}, \quad \forall j = 1, 2, \ldots, n_c. \tag{13}$$

Regarding the above-mentioned assumptions, Assumptions 1 and 4 are similar to the ones used in [19], while we make a relaxed assumption on the approximation error in Assumption 3 and remove assumptions on the autonomous dynamics $\boldsymbol{f}$ and conservative inputs used in [19].

3 Safe exploration method with conservative inputs

Now, we propose the safe exploration method to guarantee the safety in the sense of satisfaction of the chance constraint (6). As shown in Fig. 1, the basic idea is to decide whether to explore or not by using the knowledge about the controlled object and disturbance. The detailed way is given as Algorithm 1 below. Here $\wedge$ is the logical conjunction, Φ is the normal cumulative distribution function,

Algorithm 1 Proposed safe exploration method

```
At every timestep $k \geq 0$, observe state $\boldsymbol{x}_k$ and generate input $\boldsymbol{u}_k$ as follows:

(i) if $\boldsymbol{x}_k \in \mathcal{X}_s \wedge \left( \left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(\eta'_k)} (d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j), \forall \delta_j \in \{\pm \bar{\delta}_j\}, \forall j = 1, \ldots, n_c \right)$
    $\boldsymbol{u}_k = \boldsymbol{\mu}(\boldsymbol{x}_k; \boldsymbol{\theta}_k) + \boldsymbol{\varepsilon}_k,$ where $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k),$
(ii) elseif $\boldsymbol{x}_k \in \mathcal{X}_s \wedge \left( \left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2 > \frac{1}{\Phi^{-1}(\eta'_k)} (d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j),$ for some $\delta_j \in \{\pm \bar{\delta}_j\} \right)$
    $\boldsymbol{u}_k = \boldsymbol{u}_k^{stay},$
(iii) else (i.e., $\boldsymbol{x}_k \notin \mathcal{X}_s$)
    $\boldsymbol{u}_k = \boldsymbol{u}_k^{back}.$
```

$$\hat{\boldsymbol{x}}_{k+1} := \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{\mu}(\boldsymbol{x}_k; \boldsymbol{\theta}_k) + \boldsymbol{\mu}_w, \qquad \eta'_k := 1 - \frac{1 - \left(\frac{\eta}{\xi^k} \right)^{\frac{1}{T}}}{n_c}, \tag{14}$$

and ξ is a positive real number that satisfies $\eta^{\frac{1}{T}} < \xi < 1$. The quantity $\hat{\boldsymbol{x}}_{k+1}$ is a one-step ahead predicted state based on the mean of the linear approximation model (9) with substitution of (8)[2]. In the case (i), the degree of exploration is adjusted by choosing the variance-covariance matrix $\boldsymbol{\Sigma}_k$ of the stochastic exploration term $\boldsymbol{\varepsilon}_k$ to satisfy the following inequality for all $j = 1, \ldots, n_c$:

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(\eta'_k)} (d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j), \forall \delta_j \in \{\pm \bar{\delta}_j\}, \tag{15}$$

where $\boldsymbol{B}' = [\boldsymbol{B}, \boldsymbol{I}]$.

Note that the case $\boldsymbol{x}_k \in \mathcal{X}_s$ (i.e., the current state satisfies all constraints) is divided to (i) and (ii) depending on the one-step ahead predicted state $\hat{\boldsymbol{x}}_{k+1}$, and we use an exploratory input only when $\left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(\eta'_k)} (d_j - \boldsymbol{h}_j^\top \hat{\boldsymbol{x}}_{k+1} - \delta_j), \forall \delta_j \in \{\pm \bar{\delta}_j\}$ holds for all j. Rough and intuitive meaning of this condition is that we

[2] Note that the means of $\boldsymbol{\varepsilon}_k$ and $\boldsymbol{w}_k$ are assumed to be $\boldsymbol{0}$ and $\boldsymbol{\mu}_w$, respectively.

allow exploration only when the next state probably stays in $\mathcal{X}_s$ even if we generate the input with $\boldsymbol{\varepsilon}_k$, given that $\boldsymbol{\Sigma}_k$ is a solution of (15).

The inputs $\boldsymbol{u}_k^{stay}$ and $\boldsymbol{u}_k^{back}$ used in the cases (ii) and (iii) are defined as below. These inputs do not contain exploring aspects, and thus we call them conservative inputs.

Definition 1. *We call $\boldsymbol{u}_k^{stay}$ a conservative input of the first kind with which $\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}$ holds if $\boldsymbol{x}_k = \boldsymbol{x} \in \mathcal{X}_s$ occurs at timestep $k \geq 0$.*

Definition 2. *We call $\boldsymbol{u}_k^{back}$, $\boldsymbol{u}_{k+1}^{back}$, $\ldots$, $\boldsymbol{u}_{k+\tau-1}^{back}$ a sequence of conservative inputs of the second kind with which for some $j \leq \tau$, $\Pr\{\boldsymbol{x}_{k+j} \in \mathcal{X}_s\} \geq \xi$ holds if $\boldsymbol{x}_k = \boldsymbol{x} \notin \mathcal{X}_s$ occurs at timestep $k \geq 1$. That is, using these inputs in this order, the state moves back to $\mathcal{X}_s$ within τ steps with a probability of at least ξ.*

We give sufficient conditions to construct these $\boldsymbol{u}_k^{stay}$ and $\boldsymbol{u}_k^{back}$ in Section 4.3. As shown in the examples in Section 5.1, the controllability index of the linear approximation model can be used as a clue to find the positive integer τ.

Figure 2 illustrates how the proposed method switches the inputs differently in accordance with the three cases. In the case (i), the state constraints are satisfied and the input contains exploring aspect, (ii) the state constraints are satisfied but the input does not contain exploring aspect, and (iii) the state constraints are not satisfied and the input does not contain exploring aspect.

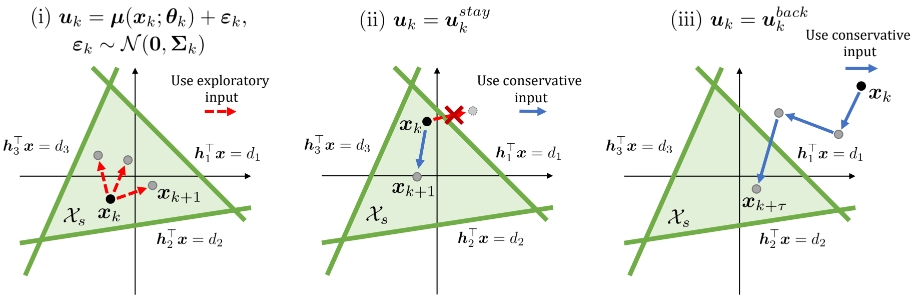
(i) $\boldsymbol{u}_k = \boldsymbol{\mu}(\boldsymbol{x}_k;\boldsymbol{\theta}_k) + \boldsymbol{\varepsilon}_k$,
$\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k)$
Use exploratory input
$\boldsymbol{h}_3^\top \boldsymbol{x} = d_3$
$\boldsymbol{h}_1^\top \boldsymbol{x} = d_1$
$\boldsymbol{x}_{k+1}$
$\mathcal{X}_s$
$\boldsymbol{x}_k$
$\boldsymbol{h}_2^\top \boldsymbol{x} = d_2$
(ii) $\boldsymbol{u}_k = \boldsymbol{u}_k^{stay}$
Use conservative input
$\boldsymbol{x}_k$
$\boldsymbol{h}_3^\top \boldsymbol{x} = d_3$
$\boldsymbol{h}_1^\top \boldsymbol{x} = d_1$
$\boldsymbol{x}_{k+1}$
$\mathcal{X}_s$
$\boldsymbol{h}_2^\top \boldsymbol{x} = d_2$
(iii) $\boldsymbol{u}_k = \boldsymbol{u}_k^{back}$
Use conservative input
$\boldsymbol{x}_k$
$\boldsymbol{h}_3^\top \boldsymbol{x} = d_3$
$\boldsymbol{h}_1^\top \boldsymbol{x} = d_1$
$\boldsymbol{x}_{k+\tau}$
$\mathcal{X}_s$
$\boldsymbol{h}_2^\top \boldsymbol{x} = d_2$

Fig. 2. Illustration of the proposed method for a case of $n = 2$ and $n_c = 3$. The proposed method switches two types of inputs in accordance with the current and one-step ahead predicted state information: exploratory inputs generated by a deterministic base policy and a Gaussian exploration term are used in the case (i), while the conservative ones that do not contain exploring aspect are used in the cases (ii) and (iii).

The proposed method, Algorithm 1, switches the exploratory inputs and the conservative ones in accordance with the current and one-step ahead predicted state information by using prior knowledge of both the controlled object and disturbance, while the previous work [19] only used that of the controlled object.

In addition, this method adjusts the degree of exploration to an appropriate level by restricting $\boldsymbol{\Sigma}_k$ of the exploration term $\boldsymbol{\varepsilon}_k$ to a solution of (15), which also contains prior knowledge of both the controlled object and disturbance.

4 Theoretical guarantee for chance constraint satisfaction

In this section, we provide theoretical results regarding the safe exploration method we introduced in the previous section. In particular, we theoretically prove that the proposed method makes the state constraints satisfied with a pre-specified probability, i.e., makes the chance constraint (6) hold, at every timestep.

We consider the case (i) in Algorithm 1 in Subsection 4.1 and the case (iii) in Subsection 4.2, respectively. We provide Theorem 1 regarding the construction of conservative inputs used in the cases (ii) and (iii) in Subsection 4.3. Then, in Subsection 4.4, we provide Theorem 2, which shows that the proposed method makes the chance constraint (6) satisfied at every timestep k under Assumptions 1–4. Proofs of the lemmas and theorems described in this section are given in Appendix B.

4.1 Theoretical result on the exploratory inputs generated with a deterministic base policy and a Gaussian exploration term

First, we consider the case (i) in Algorithm 1 in which we generate an input containing exploring aspect according to (8) with a deterministic base policy and a Gaussian exploration term. The following lemma holds.

Lemma 1. *Let $q \in (0.5,\ 1)$. Suppose Assumptions 1, 2, and 3 hold. Generate input $\boldsymbol{u}_k$ according to (8) when the state of the nonlinear system (1) at timestep k is $\boldsymbol{x}_k$. Then, the following inequality is a sufficient condition for $\Pr\{\boldsymbol{h}_j^\top \boldsymbol{x}_{k+1} \leq d_j\} \geq q,\ \forall j = 1, \ldots, n_c$:*

$$\left\| \boldsymbol{h}_j^\top \boldsymbol{B}' \begin{bmatrix} \boldsymbol{\Sigma}_k & \\ & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2 \leq \frac{1}{\Phi^{-1}(q)} \left\{ d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{\mu}(\boldsymbol{x}_k; \boldsymbol{\theta}_k) + \boldsymbol{\mu}_w) + \delta_j \right\},$$
$$\forall j = 1,\ 2,\ \ldots,\ n_c, \quad \forall \delta_j \in \{\bar{\delta}_j,\ -\bar{\delta}_j\}, \tag{16}$$

where $\boldsymbol{B}' = [\boldsymbol{B}, \boldsymbol{I}]$ and Φ is the normal cumulative distribution function.

Proof is given in Appendix B.1. This lemma is proved with the equivalent transformation of a chance constraint into its deterministic counterpart [5, §4.4.2] and holds since the disturbance $\boldsymbol{w}_k$ follows a normal distribution and is uncorrelated to the input $\boldsymbol{u}_k$ according to Assumption 2 and (8). Furthermore, this lemma shows that, in the case (i), the state satisfies the constraints with an arbitrary probability $q \in (0.5, 1)$ by adjusting the variance-covariance matrix $\boldsymbol{\Sigma}_k$ used to generate the Gaussian exploration term $\boldsymbol{\varepsilon}_k$ so that the inequality (16) would be satisfied.

4.2 Theoretical result on the conservative inputs of the second kind

Next, we consider the case (iii) in Algorithm 1 in which the state constraints are not satisfied. In this case, we use the conservative inputs of the second kind defined in Definition 2. Regarding this situation, the following lemma holds.

Lemma 2. *Suppose we use input sequence $\boldsymbol{u}_k^{back}$, $\boldsymbol{u}_{k+1}^{back}$, $\ldots$, $\boldsymbol{u}_{k+j-1}^{back}$ $(j < \tau)$ given in Definition 2 when $\boldsymbol{x}_{k-1} \in \mathcal{X}_s$ and $\boldsymbol{x}_k = \boldsymbol{x} \notin \mathcal{X}_s$ occur. Also suppose $\boldsymbol{x}_k \in \mathcal{X}_s \Rightarrow \Pr\{\boldsymbol{x}_{k+1} \in \mathcal{X}_s\} \geq p$ holds with $p \in (0,1)$. Then $\Pr\{\boldsymbol{x}_k \in \mathcal{X}_s\} \geq \xi^k p^\tau$ holds for all $k = 1, 2, \ldots, T$ if $\boldsymbol{x}_0 \in \mathcal{X}_s$.*

Proof is given in Appendix B.2. This lemma gives us a theoretical guarantee to make a state violating the constraints satisfy them with a desired probability after a certain number of timesteps if we use conservative inputs (or input sequence) defined in Definition 2.

4.3 Theoretical result on how to generate conservative inputs

As shown in Algorithm 1, our proposed method uses conservative inputs $\boldsymbol{u}_k^{stay}$ and $\boldsymbol{u}_k^{back}$ given in Definitions 1 and 2, respectively. Therefore, when we try to apply this method to real problems, we need to construct such conservative inputs. To address this issue, in this subsection, we introduce sufficient conditions to construct those conservative inputs, which are given by using prior knowledge of the controlled object and disturbance. Namely, regarding $\boldsymbol{u}_k^{stay}$ and $\boldsymbol{u}_k^{back}$ used in Algorithm 1, we have the following theorem.

Theorem 1. *Let $q \in (0.5, 1)$. Suppose Assumptions 1, 2 and 3 hold. Then, if input $\boldsymbol{u}_k$ satisfies the following inequality for all $j = 1, 2, \ldots, n_c$ and $\delta_j \in \{\bar{\delta}_j, -\bar{\delta}_j\}$, $\Pr\{\boldsymbol{x}_{k+1} \in \mathcal{X}_s\} \geq q$ holds:*

$$d_j - \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\mu}_w) - \delta_j \geq \Phi^{-1}(q') \left\| \boldsymbol{h}_j^\top \boldsymbol{\Sigma}_w^{\frac{1}{2}} \right\|_2, \tag{17}$$

where $q' = 1 - \frac{1-q}{n_c}$.

In addition, if input sequence $\boldsymbol{U}_k = [\boldsymbol{u}_k^\top, \boldsymbol{u}_{k+1}^\top, \ldots, \boldsymbol{u}_{k+\tau-1}^\top]^\top$ satisfies the following inequality for all $j = 1, 2, \ldots, n_c$ and $\Delta_j \in \{\bar{\Delta}_j, -\bar{\Delta}_j\}$, $\Pr\{\boldsymbol{x}_{k+\tau} \in \mathcal{X}_s\} \geq q$ holds:

$$d_j - \boldsymbol{h}_j^\top \left(\boldsymbol{A}^\tau \boldsymbol{x}_k + \hat{\boldsymbol{B}}\boldsymbol{U}_k + \hat{\boldsymbol{C}}\hat{\boldsymbol{\mu}}_w \right) - \Delta_j \geq \Phi^{-1}(q') \left\| \boldsymbol{h}_j^\top \hat{\boldsymbol{C}} \begin{bmatrix} \boldsymbol{\Sigma}_w & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_w \end{bmatrix}^{\frac{1}{2}} \right\|_2, \tag{18}$$

where $\hat{\boldsymbol{\mu}}_w = \left[\boldsymbol{\mu}_w^\top, \ldots, \boldsymbol{\mu}_w^\top\right]^\top \in \mathbb{R}^{n\tau}$, $\hat{\boldsymbol{B}} = [\boldsymbol{A}^{\tau-1}\boldsymbol{B}, \boldsymbol{A}^{\tau-2}\boldsymbol{B}, \ldots, \boldsymbol{B}]$ and $\hat{\boldsymbol{C}} = [\boldsymbol{A}^{\tau-1}, \boldsymbol{A}^{\tau-2}, \ldots, \boldsymbol{I}]$.

Sketch of Proof. First, from Bonferroni's inequality, the following relation holds for $q' = 1 - \frac{1-q}{n_c}$ and $\forall \delta_j$, $\forall j = 1, \ldots, n_c$:

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \geq q \Leftarrow \Pr\left\{ \boldsymbol{h}_j^\top (\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k) + \delta_j \leq d_j \right\} \geq q'. \tag{19}$$

Next, as input $\boldsymbol{u}_k$ and disturbance $\boldsymbol{w}_k$ follow normal distributions and are uncorrelated (Assumption 2 and (8)), the following relation holds [5, §4.4.2]:

$$\begin{aligned} &\Pr\{\boldsymbol{h}_j^\top(\boldsymbol{A}\boldsymbol{x}_k+\boldsymbol{B}\boldsymbol{u}_k+\boldsymbol{w}_k)+\delta_j \le d_j\} \ge q' \\ &\Leftrightarrow d_j-\boldsymbol{h}_j^\top(\boldsymbol{A}\boldsymbol{x}_k+\boldsymbol{B}\boldsymbol{u}_k)-\delta_j-\boldsymbol{h}_j^\top\boldsymbol{\mu}_w \ge \Phi^{-1}(q')\left\|\boldsymbol{h}_j^\top\boldsymbol{\Sigma}_w^{\frac{1}{2}}\right\|_2. \end{aligned} \tag{20}$$

Therefore, the first part of the theorem is proved. The second part of the theorem is proved in the same way. Full proof is given in Appendix B.3. □

This theorem means that, if we find solutions of (17) with $q'=1-\frac{1-\left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}}{n_c}$ and (18) with $q'=1-\frac{1-\xi}{n_c}$, they can be used as the conservative inputs $\boldsymbol{u}_k^{stay}$ and $\boldsymbol{u}_k^{back}$ in Definitions 1 and 2, respectively. Since (17) and (18) are linear w.r.t. $\boldsymbol{u}_k$ and $\boldsymbol{U}_k$, we can use linear programming solvers to find the solutions. Concrete examples of the conditions given in this theorem are shown in simulation evaluations in Section 5.

4.4 Main theoretical result: Theoretical guarantee for chance constraint satisfaction

Using the complementary theoretical results described so far, we show our main theorem that guarantees the satisfaction of the safety when we use our proposed safe exploration method, Algorithm 1, even with the existence of disturbance.

Theorem 2. *Let $\eta \in (0.5, 1)$. Suppose Assumptions 1 through 4 hold. Then, by generating input $\boldsymbol{u}_k$ according to Algorithm 1, chance constraint (6) are satisfied at every timestep $k=1,2,\ldots,T$.*

Sketch of Proof. First, consider the case of (i) in Algorithm 1. From Lemma 1, Assumptions 3 and 4, and Bonferroni's inequality,

$$\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \ge \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}} \tag{21}$$

holds if the input $\boldsymbol{u}_k$ is generated by (8) with $\boldsymbol{\Sigma}_k$ satisfying (15), and thus, chance constraint (6) is satisfied for $k=1,2,\ldots,T$.

Next, in the case of (ii) in Algorithm 1, by generating an input as $\boldsymbol{u}_k=\boldsymbol{u}_k^{stay}$ that is defined in Definition 1, $\Pr\{\boldsymbol{H}\boldsymbol{x}_{k+1} \preceq \boldsymbol{d}\} \ge \left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}}$ holds when $\boldsymbol{x}_k \in \mathcal{X}_s$.

Finally, by generating input as $\boldsymbol{u}_k=\boldsymbol{u}_k^{back}$ in case (iii) of Algorithm 1, $\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \ge \eta$ holds for any $\boldsymbol{x}_k \in \mathbb{R}^n$, $k=1,2,\ldots,T$ from Lemma 2. Hence, noting $\left(\frac{\eta}{\xi^k}\right)^{\frac{1}{\tau}} > \eta$, $\Pr\{\boldsymbol{H}\boldsymbol{x}_k \preceq \boldsymbol{d}\} \ge \eta$ is satisfied for $k=1,2,\ldots,T$. Full proof is given in Appendix B.4. □

The theoretical guarantee of safety proved in Theorem 2 is obtained with the equivalent transformation of a chance constraint into its deterministic counterpart under the assumption on disturbances (Assumption 4). That is, this theoretical result holds since the disturbance follows a normal distribution and is uncorrelated to the input. The proposed method, however, can be applicable to deal with other types of disturbance if the sufficient part holds with a certain transformation.

5 Simulation evaluation

5.1 Simulation conditions

We evaluated the validity of the proposed method with the inverted-pendulum provided as “Pendulum-v0” in OpenAI Gym [6] and the four-bar parallel link robot manipulator with two degrees of freedom dealt in [18]. Configuration figures of both problems are provided in Fig. C.1 in Appendix. We added external disturbances to these problems.

Inverted-pendulum: The discrete-time dynamics of this problem is given by

$$\begin{aligned}\begin{bmatrix}\phi_{k+1}\\ \zeta_{k+1}\end{bmatrix} &= \begin{bmatrix}\phi_k + T_s\zeta_k\\ \zeta_k - T_s\frac{3g}{2\ell}\sin(\phi_k+\pi)\end{bmatrix} + \begin{bmatrix}0\\ T_s\frac{3}{m\ell^2}\end{bmatrix}u_k + \boldsymbol{w}_k\\ &=: \boldsymbol{f}(\boldsymbol{x}_k) + \boldsymbol{G}u_k + \boldsymbol{w}_k, \qquad (22)\end{aligned}$$

where $\phi_k \in \mathbb{R}$ and $\zeta_k \in \mathbb{R}$ are the angle and rotating speed of the pendulum and $\boldsymbol{x}_k = [\phi_k, \zeta_k]^\top$. Further, $u_k \in \mathbb{R}$ is an input torque, T_s is a sampling period, and $\boldsymbol{w}_k \in \mathbb{R}^2$ is the disturbance where $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, $\boldsymbol{\mu}_w = [\mu_{w,\phi}, \mu_{w,\zeta}]^\top \in \mathbb{R}^2$ and $\boldsymbol{\Sigma}_w = \mathrm{diag}(\sigma^2_{w,\phi}, \sigma^2_{w,\zeta}) \in \mathbb{R}^{2\times 2}$. Concrete values of these and the other variables used in this evaluation are listed in Table C.1 in Appendix. We use the following linear approximation model of the above nonlinear system:

$$\begin{aligned}\boldsymbol{x}_{k+1} &\simeq \begin{bmatrix}1 & T_s\\ 0 & 1\end{bmatrix}\boldsymbol{x}_k + \begin{bmatrix}0\\ T_s\frac{3}{m\ell^2}\end{bmatrix}u_k + \boldsymbol{w}_k\\ &=: \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}u_k + \boldsymbol{w}_k. \qquad (23)\end{aligned}$$

The approximation errors $\boldsymbol{e}$ in (10) is given by

$$\boldsymbol{e}(\boldsymbol{x}, u) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{G}u - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}u) = \begin{bmatrix}0\\ -T_s\frac{3g}{2\ell}\sin(\phi+\pi)\end{bmatrix}. \qquad (24)$$

We set constraints on ζ_k as $-6 \leq \zeta_k \leq 6$, $\forall k = 1, \ldots, T$. This condition becomes

$$\boldsymbol{h}_1^\top \boldsymbol{x}_k \leq d_1,\ \boldsymbol{h}_2^\top \boldsymbol{x}_k \leq d_2,\ \forall k = 1, \ldots, T, \qquad (25)$$

where $\boldsymbol{h}_1^\top = [0, 1]$, $\boldsymbol{h}_2^\top = [0, -1]$, $d_1 = d_2 = 6$, and $n_c = 2$. Therefore, Assumption 4 holds since $\boldsymbol{h}_j^\top \boldsymbol{B} \neq 0$, $j \in \{1, 2\}$. Furthermore, the approximation model given by (23) is controllable because of its coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and its controllability index is 2. According to this result, we set $\tau = 2$ and we have

$$\sup_{\boldsymbol{x}\in\mathbb{R}^2,\ u\in\mathbb{R}} |\boldsymbol{h}_j^\top \boldsymbol{e}(\boldsymbol{x}, u)| = T_s\frac{3g}{2\ell},\quad j \in \{1, 2\}, \qquad (26)$$

$$\sup_{\boldsymbol{x}\in\mathbb{R}^2,\ u\in\mathbb{R}} |\boldsymbol{h}_j^\top (\boldsymbol{A} + \boldsymbol{I})\boldsymbol{e}(\boldsymbol{x}, u)| = T_s\frac{3g}{\ell},\quad j \in \{1, 2\}, \qquad (27)$$

since $|\sin(\phi + \pi)| \leq 1$, $\forall \phi \in \mathbb{R}$. Therefore we used in this evaluation $T_s\frac{3g}{2\ell}$ and $T_s\frac{3g}{\ell}$ as $\bar{\delta}_j$ and $\bar{\Delta}_j$, respectively, and they satisfy Assumption 3.

Regarding immediate cost, we let

$$c_{k+1} = \Big(\{(\phi_k + \pi) \bmod 2\pi\} - \pi\Big)^2 + 0.1\zeta_k^2 + 0.001u_k^2. \tag{28}$$

The first term corresponds to swinging up the pendulum and keeping it inverted. Furthermore, in our method, we used the following conservative inputs:

$$u_k^{stay} = -\frac{m\ell^2}{3T_s}(\zeta_k + \mu_{w,\phi}),\ \begin{bmatrix} u_k^{back} \\ u_{k+1}^{back} \end{bmatrix} = \begin{bmatrix} -\frac{m\ell^2}{3T_s}(\zeta_k + 2\mu_{w,\phi}) \\ 0 \end{bmatrix}. \tag{29}$$

Both of these inputs satisfy the inequalities in Theorem 1 with the parameters in Table C.1, and can be used as conservative inputs defined in Definitions 1 and 2.

Four-bar parallel link robot manipulator: We let $\boldsymbol{x} = [q_1,\ q_2,\ \varpi_1,\ \varpi_2]^\top \in \mathbb{R}^4$ and $\boldsymbol{u} = [v_1,\ v_2]^\top \in \mathbb{R}^2$ where q_1, q_2 are angles of links of a robot, ϖ_1, ϖ_2 are their rotating speed and v_1, v_2 are armature voltages from an actuator. The discrete-time dynamics of a robot manipulator with an actuator including external disturbance $\boldsymbol{w}_k \in \mathbb{R}^4$ where $\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, $\boldsymbol{\mu}_w = [\mu_{w,q_1}, \mu_{w,q_2}, \mu_{w,\varpi_1}, \mu_{w,\varpi_2}]^\top \in \mathbb{R}^4$ and $\boldsymbol{\Sigma}_w = \mathrm{diag}(\sigma_{w,q_1}^2, \sigma_{w,q_2}^2, \sigma_{w,\varpi_1}^2, \sigma_{w,\varpi_2}^2) \in \mathbb{R}^{4\times4}$ is given by

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \begin{bmatrix} q_{1_k} + T_s\varpi_{1_k} \\ q_{2_k} + T_s\varpi_{2_k} \\ \varpi_{1_k} - T_s\frac{\hat{d}_{11}}{\hat{m}_{11}}\varpi_{1_k} - T_s\frac{V_1}{\hat{m}_{11}}\cos q_{1_k} \\ \varpi_{2_k} - T_s\frac{\hat{d}_{22}}{\hat{m}_{22}}\varpi_{2_k} - T_s\frac{V_2}{\hat{m}_{22}}\cos q_{2_k} \end{bmatrix} + T_s\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{\alpha}{\hat{m}_{11}} & 0 \\ 0 & \frac{\alpha}{\hat{m}_{22}} \end{bmatrix}\boldsymbol{u}_k + \boldsymbol{w}_k \\ &=: \boldsymbol{f}(\boldsymbol{x}_k) + \boldsymbol{G}\boldsymbol{u}_k + \boldsymbol{w}_k, \end{aligned} \tag{30}$$

where $\hat{m}_{ii} = \eta^2 J_{mi} + M_{ii}, \hat{d}_{ii} = \eta^2\left(D_{mi} + \frac{K_tK_b}{R}\right), i \in \{1,2\}, \alpha = \frac{\eta K_a K_t}{R}$. The definitions of symbols in (30) and their specific values except the sampling period T_s are given in [18]. Derivation of (30) is detailed in Appendix C.2. Similarly, we obtain the following linear approximation model of (30) by ignoring gravity term:

$$\begin{aligned}\boldsymbol{x}_{k+1} &\simeq \begin{bmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & (1 - T_s\frac{\hat{d}_{11}}{\hat{m}_{11}}) & 0 \\ 0 & 0 & 0 & (1 - T_s\frac{\hat{d}_{22}}{\hat{m}_{22}}) \end{bmatrix}\begin{bmatrix} q_{1_k} \\ q_{2_k} \\ \varpi_{1_k} \\ \varpi_{2_k} \end{bmatrix} + T_s\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{\alpha}{\hat{m}_{11}} & 0 \\ 0 & \frac{\alpha}{\hat{m}_{22}} \end{bmatrix}\boldsymbol{u}_k + \boldsymbol{w}_k \\ &=: \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k. \end{aligned} \tag{31}$$

In the same way as the setting of the inverted pendulum problem described above, we set constraints on the upper and lower bounds regarding rotating speed ϖ_1 and ϖ_2 with $\boldsymbol{h}_1 = [0,0,1,0]^\top$, $\boldsymbol{h}_2 = [0,0,-1,0]^\top$, $\boldsymbol{h}_3 = [0,0,0,1]^\top$, $\boldsymbol{h}_4 = [0,0,0,-1]^\top$. Since $|\cos q_i| \leq 1,\ i \in \{1,2\}$, we have the following relations:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^4, \boldsymbol{u}\in\mathbb{R}^2} |\boldsymbol{h}_j^\top \boldsymbol{e}(\boldsymbol{x},\boldsymbol{u})| = \begin{cases} T_s\frac{V_1}{\hat{m}_{11}}, & j \in \{1,2\} \\ T_s\frac{V_2}{\hat{m}_{22}}, & j \in \{3,4\} \end{cases}, \tag{32}$$

$$\sup_{\boldsymbol{x}\in\mathbb{R}^4, \boldsymbol{u}\in\mathbb{R}^2} |\boldsymbol{h}_j^\top (\boldsymbol{A}+\boldsymbol{I})\boldsymbol{e}(\boldsymbol{x},\boldsymbol{u})| = \begin{cases} |2 - T_s\frac{\hat{d}_{11}}{\hat{m}_{11}}|T_s\frac{V_1}{\hat{m}_{11}}, & j \in \{1,2\} \\ |2 - T_s\frac{\hat{d}_{22}}{\hat{m}_{22}}|T_s\frac{V_2}{\hat{m}_{22}}, & j \in \{3,4\} \end{cases}. \tag{33}$$

We use them as $\bar{\delta}_j$ and $\bar{\varDelta}_j$, and therefore Assumption 3 holds. Assumption 4 also holds with $\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4$ and $\boldsymbol{B}$. In this setting, we used immediate cost

$$\begin{aligned} c_{k+1} = {} & 2\Big(\{(q_{1_k}+\pi) \bmod 2\pi\} - \pi\Big)^2 + 2\Big(\{((q_{2_k}+\pi) - 5\pi/6) \bmod 2\pi\} - \pi\Big)^2 \\ & + 0.1(\varpi_{1_k}^2 + \varpi_{2_k}^2) + 0.001\boldsymbol{u}_k^\top \boldsymbol{u}_k. \end{aligned} \tag{34}$$

The first two terms corresponds to changing the pose of manipulator to the one depicted on the right in Fig. C.2 in Appendix and keeping that pose. Furthermore, in our method, we used the following conservative inputs:

$$\boldsymbol{u}_k^{stay} = \begin{bmatrix} -\frac{1}{b_1}\{(1-a_1)\varpi_{1_k} + (1-a_1)\mu_{w,\varpi_1}\} \\ -\frac{1}{b_2}\{(1-a_2)\varpi_{2_k} + (1-a_2)\mu_{w,\varpi_2}\} \end{bmatrix}, \tag{35}$$

$$\boldsymbol{u}_k^{back} = \begin{bmatrix} -\frac{1}{(1-a_1)b_1}\{(1-a_1)^2\varpi_{1_k} + (2-a_1)\mu_{w,\varpi_1}\} \\ -\frac{1}{(1-a_2)b_2}\{(1-a_2)^2\varpi_{2_k} + (2-a_2)\mu_{w,\varpi_2}\} \end{bmatrix}, \ \boldsymbol{u}_{k+1}^{back} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{36}$$

where a_1, a_2, b_1 and b_2 are derived from elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ and they are $a_1 = T_s\hat{d}_{11}/\hat{m}_{11}$, $a_2 = T_s\hat{d}_{22}/\hat{m}_{22}$, $b_1 = T_s\alpha/\hat{m}_{11}$, and $b_2 = T_s\alpha/\hat{m}_{22}$. Both of these inputs satisfy the inequalities in Theorem 1 with the parameters in Table C.2, and thus, they can be used as conservative inputs defined in Definitions 1 and 2.

Reinforcement learning algorithm and reference method: We have combined our proposed safe exploration method (Algorithm 1) with the Deep Deterministic Policy Gradient (DDPG) algorithm [16], a representative RL algorithm applicable to (7) and (8), in each experimental setting with the immediate costs and conservative inputs described above. We also combined the safe exploration method given in the previous work [19] that does not take disturbance into account with the DDPG algorithm for the reference where we set $u_k^{stay} = 0$ as in the original paper. The network structure and hyperparameters we used throughout this evaluation are listed in Tables C.1 and C.2 in Appendix.

Parameters for safe exploration: We set the pre-specified probabilities in both problems to be $\eta = 0.95$. Other parameters for safe exploration are listed in Tables C.1 and C.2 in Appendix.

5.2 Simulation results

We evaluated our method and the previous one with 100 episodes × 10 runs of the simulation (each episode consists of 100 timesteps). The source code is publicly available as described in Code Availability Statement. The computational resource and running time information for this evaluation is given in Appendix C.4.

Figure 3 shows the results of the cumulative costs at each episode and the relative frequencies of constraint satisfaction at each timestep. The lines shown in the left figures are the mean values of the cumulative cost at each episode calculated over the 10 runs, while the shaded areas show their 95% confidence intervals. We can see that both methods enabled to reduce their cumulative costs

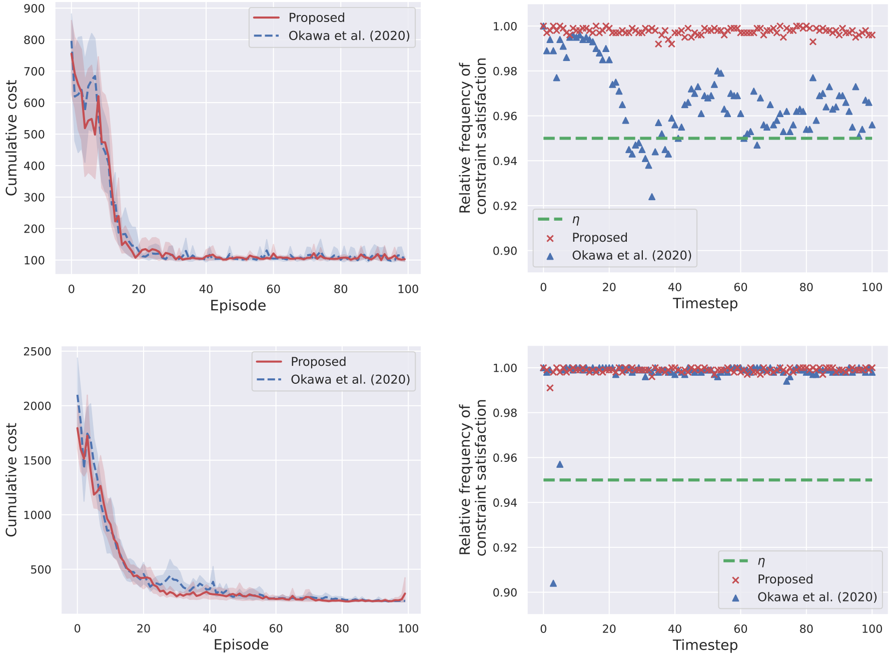

Fig. 3. Simulation results with (**Top**) an inverted-pendulum and (**Bottom**) a four-bar parallel link robot manipulator: (**Left**) Cumulative costs at each episode, (**Right**) Relative frequencies of constraint satisfaction at each timestep. Both the proposed method (red) and the previous one (blue) [19] enabled to reduce their cumulative costs; however, only the proposed method made the relative frequencies of constraint satisfaction greater than or equal to η for all timesteps in both experimental settings.

as the number of episode increases. However, as shown in the right figures, the previous method [19] (blue triangles) could not meet the chance constraint (6) (went below the green dashed lines that show the pre-specified probability η) at several timesteps. In contrast, our proposed method (red crosses) could make the relative frequencies of constraint satisfaction greater than or equal to η for all timesteps. Both simulations support our theoretical results and show the effectiveness of the proposed method.

6 Limitations

There are two main things we need to care about to use the proposed method. First, although it is relaxed compared to the previous work [19], the controlled object and disturbance should satisfy several conditions and we need partial prior knowledge about them as described in Assumptions 1 through 4. In addition, the proposed method requires calculations including matrices, vectors, nonlinear functions and probabilities. This additional computational cost may become a problem if the controller should be implemented as an embedded system.

7 Conclusion

In this study, we proposed a safe exploration method for RL to guarantee the safety during learning under the existence of disturbance. The proposed method uses partial prior knowledge of both the controlled object and disturbances. We theoretically proved that the proposed method achieves the satisfaction of explicit state constraints with a pre-specified probability at every timestep even when the controlled object is exposed to the disturbance following a normal distribution. Sufficient conditions to construct conservative inputs used in the proposed method are also provided for its implementation. We also experimentally showed the validity and effectiveness of the proposed method through simulation evaluation using an inverted pendulum and a four-bar parallel link robot manipulator. Our future work includes the application of the proposed method to real environments.

Code Availability Statement The source code to reproduce the results of this study is available at https://github.com/FujitsuResearch/SafeExploration

Acknowledgements The authors thank Yusuke Kato for the fruitful discussions on theoretical results about the proposed method. The authors also thank anonymous reviewers for their valuable feedback. This work has been partially supported by Fujitsu Laboratories Ltd and JSPS KAKENHI Grant Number JP22H01513.

References

1. Achiam, J., Held, D., Tamar, A., Abbeel, P.: Constrained policy optimization. In: Proc. of the 34th International Conference on Machine Learning. pp. 22–31 (2017)
2. Ames, A.D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., Tabuada, P.: Control barrier functions: Theory and applications. In: Proc. of the 18th European Control Conference. pp. 3420–3431 (2019)
3. Berkenkamp, F., Turchetta, M., Schoellig, A., Krause, A.: Safe model-based reinforcement learning with stability guarantees. In: Advances in Neural Information Processing Systems 30. pp. 908–919 (2017)
4. Biyik, E., Margoliash, J., Alimo, S.R., Sadigh, D.: Efficient and safe exploration in deterministic Markov decision processes with unknown transition models. In: Proc. of the 2019 American Control Conference. pp. 1792–1799 (2019)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
6. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: OpenAI Gym. arXiv preprint, arXiv:1606.01540 (2016), the code is available at https://github.com/openai/gym with the MIT License
7. Cheng, R., Orosz, G., Murray, R.M., Burdick, J.W.: End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In: Proc. of the 33rd AAAI Conference on Artificial Intelligence. pp. 3387–3395 (2019)
8. Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., Ghavamzadeh, M.: Lyapunov-based safe policy optimization for continuous control. arXiv preprint, arXiv:1901.10031 (2019)

9. Fan, D.D., Nguyen, J., Thakker, R., Alatur, N., Agha-mohammadi, A.a., Theodorou, E.A.: Bayesian learning-based adaptive control for safety critical systems. In: Proc. of the 2020 IEEE International Conference on Robotics and Automation. pp. 4093–4099 (2020)
10. García, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research **16**, 1437–1480 (2015)
11. Ge, Y., Zhu, F., Ling, X., Liu, Q.: Safe Q-learning method based on constrained Markov decision processes. IEEE Access **7**, 165007–165017 (2019)
12. Glavic, M., Fonteneau, R., Ernst, D.: Reinforcement learning for electric power system decision and control: Past considerations and perspectives. In: Proc. of the 20th IFAC World Congress. pp. 6918–6927 (2017)
13. Khojasteh, M.J., Dhiman, V., Franceschetti, M., Atanasov, N.: Probabilistic safety constraints for learned high relative degree system dynamics. In: Proc. of the 2nd Conference on Learning for Dynamics and Control. pp. 781–792 (2020)
14. Kiran, B.R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A.A.A., Yogamani, S., Pérez, P.: Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems (2021), (Early Access)
15. Koller, T., Berkenkamp, F., Turchetta, M., Boedecker, J., Krause, A.: Learning-based model predictive control for safe exploration and reinforcement learning. arXiv preprint, arXiv:1906.12189 (2019)
16. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint, arXiv:1509.02971 (2015)
17. Liu, Z., Zhou, H., Chen, B., Zhong, S., Hebert, M., Zhao, D.: Safe model-based reinforcement learning with robust cross-entropy method. ICLR 2021 Workshop on Security and Safety in Machine Learning Systems (2021)
18. Namerikawa, T., Matsumura, F., Fujita, M.: Robust trajectory following for an uncertain robot manipulator using H_∞ synthesis. In: Proc. of the 3rd European Control Conference. pp. 3474–3479 (1995)
19. Okawa, Y., Sasaki, T., Iwane, H.: Automatic exploration process adjustment for safe reinforcement learning with joint chance constraint satisfaction. In: Proc. of the 21st IFAC World Congress. pp. 1588–1595 (2020)
20. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, 2nd edn. (2018)
21. Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J.E., Ibarz, J., Finn, C., Goldberg, K.: Recovery RL: Safe reinforcement learning with learned recovery zones. IEEE Robotics and Automation Letters **6**(3), 4915–4922 (2021)
22. Yang, T.Y., Rosca, J., Narasimhan, K., Ramadge, P.J.: Projection-based constrained policy optimization. In: Proc. of the 8th International Conference on Learning Representations (2020)